



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

교육학석사학위논문

Analyzing the Cohesion of the Reading
Passages on Standardized English Tests:
CSAT, TEPS, EIKEN and TOEFL

표준화 영어시험 독해지문의 응집성 분석:
대학수학능력시험, 텡스, 에이켄, 토플

2016 년 2 월

서울대학교 대학원

외국어교육과 영어전공

김 수 정

Analyzing the Cohesion of the Reading
Passages on Standardized English Tests:
CSAT, TEPS, EIKEN and TOEFL

by

SUE-JUNG KIM

A Thesis Submitted to
the Department of Foreign Language Education
in Partial Fulfillment of the Requirements
for the Degree of Master of Arts in Education

At the

Graduate School of Seoul National University

February 2016

Abstract

This thesis intends to analyze the cohesion of the reading passages on CSAT, TEPS, EIKEN, and TOEFL. Cohesion refers to the relations of meaning in a text and is realized by cohesive devices, such as references, connectives, and vocabularies. Assessing the cohesion of a text, through measuring text difficulty at a deeper level, can be considered one of the alternatives to traditional readability formulas. While a high-cohesion text usually facilitates text understanding, a low-cohesion text might hamper reading comprehension, particularly for low-knowledge readers. As the readers with low knowledge do not have sufficient knowledge to fill in the cohesion gaps, they may have a greater difficulty with building a coherent text representation. Accordingly, if the test takers' background knowledge is not the major construct of a reading test, the passages should be written cohesively so as not to seriously hinder the low-knowledge readers' understanding.

To measure the cohesion of the passages, this study used a web-based text analyzing tool, Coh-Metrix. Specifically, 11 cohesion indices relating to referential overlaps, Latent Semantic Analysis, connectives, and causal cohesion were selected from the tool. The corpus for analysis comprises a total of 373 passages, 125 from CSAT, 120 from TEPS, 75 from EIKEN, and 53 from

TOEFL.

The results of the analysis showed that there was a significant difference between the four tests for all of the indices except Incidence of All Connectives. In line with the expectation, TOEFL with the longest passages was the highest by most measures. However, for the rest of the tests, the results were mixed. The passages on TEPS were as cohesive as or more cohesive than the EIKEN passages for many indices. Moreover, the CSAT passages were the least cohesive among the tests although they were longer than the TEPS passages. The EIKEN passages were more cohesive than the CSAT passages. In addition, it was found that the cohesion of the passages on CSAT and TEPS was very inconsistent in comparison to the EIKEN and TOEFL passages. Furthermore, the difference between the tests was the largest for LSA Given/New, and also large for Noun and Stem Local Overlaps. LSA overlaps were as powerful as referential overlaps in discriminating between the tests. For causal cohesion, there was also a significant difference, though minimal. Lastly, the results and implications of this study are discussed.

Key Words: cohesion, reading assessment, standardized English test, CSAT, TEPS, EIKEN, TOEFL, Coh-Metrix

Student Number: 2003-23711

TABLE OF CONTENTS

Abstract.....	i
CHAPTER 1. INTRODUCTION	1
1.1 The Purpose of the Study	1
1.2 Research Questions	9
1.3 Organization of the Thesis	10
Chapter 2. LITERATURE REVIEW	11
2.1 Reading Comprehension	11
2.2 Reading Assessment	16
2.3 Cohesion	20
2.4 Text Length	25
2.5 Coh-Metrix	32
Chapter 3. METHODOLOGY.....	37
3.1 Corpus	37
3.2 Tool	42
3.2.1 Coh-Metrix Indices	43
3.2.2 Procedure	48
3.3 Data Analysis.....	49

Chapter 4. RESULTS AND DISCUSSION	50
4.1. Results.....	50
4.2. Discussion	59
Chapter 5. CONCLUSION	76
5.1 Major Findings.....	76
5.2 Implications	77
5.3 Limitations and Suggestions	79
REFERENCES	82
국문초록.....	96

List of Tables

Table 3.1 Cohesion Indices for Analysis	48
Table 4.1 Means and Standard Deviations for Flesch Reading Ease and Flesch-Kincaid Grade Level.....	51
Table 4.2 Means, Standard Deviations, and Welch’s ANOVA Results for Referential Cohesion	52
Table 4.3 Dunnett T3 Post hoc Results for Noun Overlap	53
Table 4.4 Dunnett T3 Post hoc Results for Argument Overlap	54
Table 4.5 Dunnett T3 Post hoc Results for Stem Overlap.....	55
Table 4.6 Means, Standard Deviations, and Welch’s ANOVA Results for Latent Semantic Analysis	56
Table 4.7 Dunnett T3 Post hoc Results for Latent Semantic Analysis ..	57
Table 4.8 Means, Standard Deviations, and Welch’s ANOVA Results for Incidence of All Connectives	58
Table 4.9 Means, Standard Deviations, and Welch’s ANOVA Results for Ratio of Causal Particles to Causal Verbs	59
Table 4.10 Dunnett T3 Post hoc Results for Causal Particles to Causal Verbs	59

List of Figures

Figure 4.1 Distribution of Stem Global Overlap for TEPS and EIKEN	
.....	72

List of Appendices

Appendix 1. CSAT Corpus.....	93
Appendix 2. TEPS Corpus	93
Appendix 3. EIKEN Corpus	94
Appendix 4. TOEFL Corpus.....	95

CHAPTER 1.

INTRODUCTION

This chapter introduces the research by presenting the motivation and the organization of the study. Section 1.1 discusses the problem and the purpose of the study. Section 1.2 presents the research questions, and the overall organization of the study is outlined in Section 1.3.

1.1 The Purpose of the Study

As a passage on a reading test is an important factor influencing test takers' performance (Alderson, 2000; Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015; Ozuru, Rowe, O'Reilly, & McNamara, 2008; Shohamy, 1984), selecting appropriate passages based on the accurate assessment of text difficulty becomes vitally important for successful reading tests. Text difficulty have often been measured by traditional readability measures, such as Flesch Reading Ease, Flesch-Kincaid Grade level, Degrees of Reading Power, and Lexile scores (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara,

Louwerse, McCarthy, & Graesser, 2010). Most readability formulas depend on the length of words and sentences in order to measure word difficulty and sentence complexity, and they have proved efficient and relevant in predicting text difficulty (McNamara, Graesser, McCarthy, & Cai, 2014). However, as conventional readability measures only focus on the superficial linguistic features, such as words and sentences, they have been criticized for ignoring the multidimensional features of text comprehension (Graesser & McNamara, 2011; McNamara et al., 2014).

Reading comprehension is more than decoding words and sentences, as it involves understanding the meaning intended by the author and also includes interpreting the text in the reader's own way. That is, the reader forms a network of meaning based on the text and, at the same time, builds the coherent text representation by utilizing his/her own past reading experiences, background knowledge, inferences and so on (Kintsch, 1998). Hence, many researchers (Clarke, 1980; Keenan, 2012; Kintsch, 2012; Leslie & Caldwell, 2009; Snyder, Caccamise, & Wise, 2005; van den Broek, 2012) argue that reading assessment must measure the multiple aspects of cognitive processes involved in reading comprehension. In the same vein, the various discourse features of texts which would influence reading comprehension, such as text organization, text genre, or text cohesion, should also be considered in assessing text difficulty.

From such multidimensional perspectives on text difficulty, cohesion

in particular is a crucial element in text comprehension. Being defined as “relations of meaning that exist within the text, and that define it as a text” (Halliday & Hasan, 1976), cohesion contributes to a better understanding of texts (Beck, McKeown, Omanson, & Pople, 1984; Beck, McKeown, Sinatra, & Loxterman, 1991; Britton & Gülgöz, 1991; McNamara, Kintsch, Songer, & Kintsch, 1996). Cohesion is realized by a variety of cohesive devices, for example, reference, conjunctions, or repetition of similar words. The texts in which ideas are well-connected are likely to be easier to understand. On the other hand, when there are cohesion gaps, a reader is forced to infer to bridge the gaps. Such inferencing would be successful for the reader with strong motivation, high-knowledge, and appropriate strategies, but would be unsuccessful for the reader with weak motivation, low-knowledge and inefficient reading skills (Kintsch, 1988; McNamara et al., 2014).

Particularly, the effect of cohesion on reading comprehension interacts with the reader’s knowledge. When confronting cohesion gaps in a low-cohesion text, low-knowledge readers do not have sufficient knowledge to repair the cohesion breaks. However, the low-knowledge readers would better understand the high-cohesion version of the text because a high-cohesion text tends to provide more textual clues, more repetitions of ideas, and clearer relations between meanings (McNamara et al., 2014; McNamara et al., 1996). Furthermore, as found by McNamara, Graesser, and Louwerse (2012), the

authors of expository texts seem to manipulate cohesion in consideration of the reader's knowledge level. For example, the authors of science texts might write the texts with greater cohesion because their readers would be unfamiliar to the scientific concepts in the texts. On the other hand, in McNamara et al. (2012), low cohesion in the social science texts can be explained by the authors' assumption that their readers would have the considerable level of relevant knowledge to understand the texts.

For foreign and second language (L2) learners, low-cohesion texts can make reading comprehension even harder. L2 readers with incomplete lexical knowledge are likely to have decoding failures in text understanding, and such failures would be more catastrophic in low-cohesion texts because these texts do not provide sufficient textual clues. Then, L2 readers would depend more on the outside-the-text resources like background knowledge. In contrast, when the texts are highly cohesive, L2 readers would be able to guess the meanings of the unfamiliar words or enhance their understanding through abundant textual information (Mehrpour & Riazi, 2004; Yano, Long, & Ross, 1994; Young, 1999).

Since increasing cohesion often involves adding connectives, repetition of vocabulary, or elaboration of information, high-cohesion texts tend to be longer than low-cohesion texts (Crossley, Louwerse, McCarthy, & McNamara, 2007; Graesser, McNamara, & Kulikowich, 2011). Obviously, short passages can be cohesive, but longer texts have the greater likelihood of

accommodating more cohesive ties (Haswell, 1988). Put differently, longer texts are not necessarily more difficult to understand than shorter texts since cohesion facilitates text understanding.

Nevertheless, there has been a common perception that longer texts are more difficult to read than shorter texts because there is more information to interpret and memorize in the longer texts. However, high cohesion would compensate for such linguistic constraints and even help create more coherent text representation by strengthening a network of ideas for the reader. In addition, many reading researchers argue that a better reading assessment should use longer passages because longer texts are more effective in inducing various aspects of reading comprehension processes than shorter texts (Keenan, 2012; Rawson & Kintsch, 2005; van den Broek, 2012). Furthermore, the authenticity argument also favors longer texts since people are likely to encounter the texts of greater length in academic and professional settings (Alderson, 2000; Crossley et al., 2007; Ozuru et al., 2008). In sum, a longer passage is recommended for reading assessment because of its cohesive tendency and its advantage in assessing the diverse facets of reading comprehension.

However, many standardized English reading tests seem to prefer shorter passages (Grasser & Hu, 2012). Brief passages are used primarily because a number of short passages are helpful to cover a wide range of topics so that the effects of reader's background knowledge would be minimized

(Alderson, 2000; Ozuru et al., 2008). In addition, some tests present only one question or two for each short passage because many questions for the same passage could lead to the problem of item interdependence (Alderson, 2000). Some English proficiency tests administered in Korea have many brief passages as well. For instance, the reading section on the Korean college entrance exam, “College Scholastic Ability Test (CSAT)” has 25 reading passages whose average word count per passage is 148. In addition, the reading section on “Test of English Proficiency developed by Seoul National University (TEPS)” comprises 40 reading passages with the average word count of 87.

On the contrary, other English proficiency tests have only a few passages with a greater length. For instance, “Test of English as a Foreign Language (TOEFL)” has three reading passages in the reading section, with the average word count of 695. The reading section on “Jitsuyo Eigo Gino Kentei, Test in Practical English Proficiency (EIKEN),” which is the most widely used English test in Japan, has five passages with the average word count of 400. In addition to the greater length, the passages on TOEFL and EIKEN are different from those on CSAT and TEPS in that each passage has a respective title. As Armbruster and Anderson (1985) pointed out, the heading provides a good introduction to the text.

Considering the cohesive tendency of a lengthy text, the differences in the average word counts between CSAT, TEPS, EIKEN, and TOEFL raise the

interesting question of whether the reading passages on the four tests are different from each other with respect to cohesion. The previous research investigated the effects of text length or the effects of simplification or elaboration on reading comprehension for L2 readers (Mehrpour & Riazi, 2004; Oh, 2001; Yano et al., 1994; Young, 1999). However, very little research assessed cohesion of the passages on reading assessment, specifically for Korean English proficiency tests. According to the previous research (Beck et al., 1984; Beck et al., 1991; Britton & Gülgöz, 1991; McNamara et al., 1996), cohesion appears to strongly influence reading comprehension. Hence, by analyzing the cohesion of the reading passages on English proficiency tests, one may predict that some passages would help text understanding whereas others would not. In particular, low-cohesion passages, which may force a reader to infer to bridge the cohesion gaps using his/her background knowledge, would be problematic for low-knowledge readers. If the test takers' background knowledge is not the major construct of the tests, the passages must be written cohesively so as not to seriously hinder the low-knowledge readers' understanding. Especially, the accurate assessment of text difficulty is very important for high-stakes exams. Thus, this study aims to investigate the cohesion of the reading passages on CSAT, TEPS, EIKEN, and TOEFL.

In an attempt to analyze text cohesion, a quantitative tool is of interest to this study. Currently, a way to analyze text cohesion qualitatively is also

available. For example, Halliday and Hasan (1976) presented the way to analyze text cohesion by codifying the text based on the cohesive categories, investigating the particular instances of cohesion, scrutinizing the actual words and phrases that consist of cohesive ties, and thus finding out what patterns of texture emerge. Generally, the qualitative ways would be richer in analysis and could readily detect the subtle features of text cohesion. However, these measures are subject to an analyst's judgment and prior knowledge (Zakaluk & Samuels, 1988). In addition, such analyses which greatly depend on human judgment are not efficient to apply at large scales. The quantitative measures, on the other hand, assess text cohesion objectively and computationally. Thus, a computational text analyzing tool, Coh-Metrix will be used to analyze text cohesion in the current study mainly because this study examines a large amount of texts. Coh-Metrix is a repository of multiple indices to analyze text, ranging from surface linguistic features to sophisticated discourse dimensions. In particular, Coh-Metrix has proven effective to measure text cohesion (McNamara et al., 2010; McNamara, Ozuru, Graesser, & Louwerse, 2006). Specifically, 11 cohesion indices, drawn from the banks of referential cohesion, Latent Semantic Analysis, connectives, and situation model measures, are selected for this study based on the previous research (Duran, Bellissens, Taylor, & McNamara, 2007; Graesser et al., 2011; Graesser et al., 2004; McNamara et al., 2010).

In sum, in order to investigate passage comprehensibility at a deeper

level, this study intends to examine the reading passages on CSAT, TEPS, EIKEN and TOEFL with respect to the cohesion indices of Coh-Metrix.

1.2 Research Questions

The present study intends to analyze the reading passages on CSAT, TEPS, EIKEN and TOEFL with respect to 11 cohesion indices of Coh-Metrix. That is, this study tries to compare cohesion between the reading passages of the four tests, specifically investigating whether the different length of the reading passages is associated with the cohesion of the texts. The shorter passages on CSAT and TEPS are predicted to be lower in cohesion than the longer passages on EIKEN and TOEFL. In addition, TOEFL with the longest passages is expected to be the most cohesive among the tests, while TEPS with the shortest passages is predicted to be the lowest. Thus, this study intends to answer the following research questions.

1. Are the reading passages on CSAT, TEPS, EIKEN and TOEFL different from each other with respect to cohesion?
2. If so, which features of cohesion are driving these differences?

1.3 Organization of the Thesis

The present study consists of five chapters. Chapter 1 introduces the motivation of the study and presents the research questions. Chapter 2 presents the literature review on reading comprehension, reading assessment, cohesion, text length and Coh-Metrix. In chapter 3, the method of the study is described in terms of the corpus, the tool, and the data analysis. Chapter 4 presents the results and discusses the research findings. Finally, Chapter 5 concludes the research with the summary of the major findings and presents the implications of the present study as well as the suggestions for future research.

Chapter 2.

LITERATURE REVIEW

This chapter presents the literature overview related to reading comprehension, reading assessment, cohesion, text length and Coh-metrix. Section 2.1 discusses reading comprehension for both first language and second language. Section 2.2 reviews research on reading assessment. Section 2.3 explains text cohesion for which the passages on reading comprehension tests are analyzed in this study. In Section 2.4, text length is discussed in relation to cohesion. Finally, Section 2.5 introduces Coh-Metrix whereby cohesion is measured.

2.1 Reading Comprehension

The extensive body of reading research has shown that reading comprehension is a complex process which can be explained in terms of the multilevel and multicomponent features (Alderson, 2000; Grabe, 2009). It is also generally acknowledged that reading comprises lower-level and higher-level processes (Alderson et al., 2015; Grabe, 2009). The lower level processes involve word recognition, syntactic parsing, and encoding meaning. The higher-

level processes include a set of resources and skills, such as background knowledge, strategies, inferences, and comprehension monitoring. Both lower-level and higher-level processes operate at the same time and interact with each other as needed (Grabe, 2009).

To discuss in detail, the lower-level processes are the building blocks for successful reading comprehension. Word recognition involves a range of subskills, such as linking between letter and sound, activating semantic and syntactic resources, and accessing mental lexicon (Grabe, 2009). In addition, the reader draws information from words and sentence structure through syntactic parsing. Hence, the reader forms a semantic proposition, “a network of small packets of information linked together in a meaning unit” (Grabe, 2009, p. 31).

Next, for the higher-level processes, there are two types of text representations, a textbase model and a situation model (Kintsch, 1998). Kintsch distinguished several levels of the psychological processes associated with discourse comprehension: the surface structure, textbase, and situation model. While the surface structure deals with immediate words and syntax, both the textbase and situation model are related to the general understanding of a text. In other words, when encountering new sentences, a reader integrates the newly formed propositions into the existing network of ideas, the text model of comprehension. In order to build a coherent network of ideas, the reader adds new information or suppresses peripheral information. The successful building

of the text model is to understand the meaning intended by the author. The situation model, on the other hand, represents the reader's own interpretation of the text, which is influenced by a variety of factors including the reader's purpose, attitude, past reading experiences, background knowledge, inferences and so on. Simply put, the situation model indicates the meanings created by the reader himself/herself. The theory of text comprehension in terms of both models has proved very effective to account for some primary issues in reading research, enabling the better understanding of text features or reader resources (Grabe, 2009).

In addition, a set of skills and resources, also components of the higher-level processes, are involved in building the text model and situation model. Under the control of working memory and attentional processes, skills and resources, such as strategies, goals, inferences, background knowledge, and comprehension monitoring, play important roles for comprehension (Grabe, 2009). Put another way, a reader utilizes a variety of higher-level processing elements at hand in order to successfully construct text representation.

Generally, the processes involved in reading are similar across different languages (Grabe, 2009). All readers use language knowledge like word and syntactic information to decode meaning. They also have some specific reading purposes in mind. Furthermore, they commonly take advantage of strategies, background knowledge, or inferencing skills.

Drawing on such similarities of reading abilities across languages, it was once asserted that since first language (L1) reading abilities directly transfers to second and foreign language (L2) reading, there is little need to improve second and foreign language to enhance L2 reading (Cummins, 2000 cited in Grabe, 2009, p. 141). However, some researchers (Droop & Verhoeven, 2003) found that second and foreign language plays a critical role in L2 reading. Recently, it is believed that the effects of L1 on L2 reading are rather complex (Grabe, 2009). Different L1 languages would facilitate or interfere with L2 reading. For example, a reader whose L1 shares orthographic features with L2 is likely to recognize the words in L2 in a faster and more accurate manner. In the similar vein, all of the reading abilities are not universal. Despite some governing principles for reading comprehension, the specific operations of these principles vary across languages. For example, readers from a given L1 would use different combinations of reading strategies than those from a different L1 (Frost, 2005). Thus, recent research into L2 reading is generally interested not only in commonalities across languages but also in the effects of specific linguistic settings (Grabe, 2009).

Compared to L1 reading, a starting point for L2 readers is very different. Whereas L1 beginning readers, often children, have firmly mastered oral language, L2 beginning readers learn to speak and read at the same time (Alderson, 2000; Grabe, 2009). Because of the incomplete L2 lexicon, L2

reading is more likely to be challenging than L1 reading. For example, unlike the L1 readers who already have vast knowledge about their language, L2 readers will have the problems with word recognition or reading rate.

Accordingly, it is widely acknowledged that L2 readers need to reach a certain level of L2 proficiency, the commonly called 'language threshold' so that their L1 reading might transfer to L2 reading (Alderson, 2000; Clarke, 1980; Grabe, 2009). Due to such language restrictions, Alderson (2000) contended that L2 reading is more of language problem than of reading problem. Moreover, the language threshold is not fixed for an individual reader, but changes according to various factors including texts, reading purpose, or background knowledge.

In addition to language itself, various factors are involved in L2 reading. Readers often bring their L1 reading experiences and skills to L2 reading. Moreover, exposure to L2 reading is limited in terms of types and amount, which becomes more problematic for foreign language reading than for second language reading. In addition, readers' expectations or attitudes toward texts would vary, as the literary conventions are different across societies and cultures (Grabe, 2009).

In sum, L2 reading is as much complex as or more complex than L1 reading. L2 readers bring some abilities from L1 reading, but, at the same time, L2 readers face the challenges due to different languages and cultural backgrounds. Hence, to better understand L2 reading, the specific circumstances

in which the reader experiences should be carefully considered.

2.2 Reading assessment

Assessing reading comprehension is never an easy task not only because reading is a complex process but also because it is hidden and internal to the reader (Alderson et al., 2015). The “process” of reading would be different depending on individuals even though the “product,” what a reader has understood from a text, remains the same. For example, a test taker can get a correct answer to a comprehension question for wrong reasons, or vice versa. In addition, test takers’ responses are affected by the ways how assessment elicits such responses. Many variables, such as text features or test format, influence readers’ performance (Alderson, 2000; Alderson et al., 2015). For these reasons, it is challenging for test developers to assure that reading assessment exactly measures what it is supposed to measure.

Concerning what reading assessment should measure, there have been various opinions. First of all, the notion that reading consists of a range of subskills has influenced many test developers, thus isolating and assessing discrete skills, for example, identifying main ideas, interpreting a word, or making an inference (Alderson et al., 2015). However, such an attempt has been criticized mainly because it is not clear what sub-skills are actually being tested.

Moreover, although the students were taught such skills, they were often unable to transfer these skills to reading comprehension (Leslie & Caldwell, 2009). After all, test designers have begun to recognize that reading is more than the sum of sub-skills.

Notably, the psychological theories of reading comprehension have provided valuable insights into the constructs of reading assessment. Many researchers (Clarke, 1980; Keenan, 2012; Kintsch, 2012; Leslie & Caldwell, 2009; Snyder et al., 2005; van den Broek, 2012) asserted that reading assessment must measure the cognitive processes involved in reading comprehension. Van den Broek (2012) argued that reading assessment should evaluate a reader's skills, strategies, and cognitive structures which are involved in text representations. Kintsch (2012) also contended that psychological processes pertaining to reading must be a key component of comprehension testing. Most importantly, he argued, test-developers should be aware of the distinct levels of text representations that are assessed. That is to say, superficial understanding at the textbase model or deep comprehension at the situation model takes place depending on the tasks of the assessment. For example, the mental processes involved in recalling task may be different from those associated with conceptual sorting task.

Although a great deal of research into reading assessment has been concerned with item difficulty and measurement itself (Alderson, 2000; Mislevy

& Sabatini, 2012), the reading passages in assessment should also be subject to close scrutiny. Obviously, items and passages in a test interact with each other. However, as Shohamy (1984) argued, the passages would have a significant effect on test scores regardless of tasks or question languages.

With respect to the text, researchers have primarily focused on the factors which affect text difficulty. First of all, it is generally known that narrative texts are easier to understand than expository texts because the words and concepts from narrative texts are more familiar and concrete than those from expository texts (Alderson, 2000; McNamara et al., 2014). Next, text organization was found to influence comprehension (Carrell, 1984; Kobayashi, 2002). In Kobayashi's study (2002), more proficient students showed better performance when reading well-structured texts whereas less proficient students showed no difference. Other text features, such as cohesion and length, are also related to text difficulty, and they will be discussed in detail later in this review. In addition, a reader factor like test-takers' background knowledge was found to have a significant effect on text comprehensibility (Grabe, 2009; McNamara et al., 2014).

Since a variety of factors are involved in text difficulty, it is a great challenge for test developers to select the suitable texts for a given assessment. They often choose passages based upon readability measures. Readability measures have long been used to scale texts on difficulty, which include Flesch

Reading Ease, Flesch-Kincaid Grade level, Degrees of Reading Power, and Lexile scores. Most readability measures are dependent on two factors, word difficulty and sentence complexity (McNamara et al., 2012; Zakaluk & Samuels, 1988). Word difficulty is associated with word familiarity, frequency or length. Sentence complexity is related to sentence length or the syntactic complexity of a sentence. As difficult words or sentences tend to be long, many readability measures use the length as an important index in assessing word and sentence difficulty.

Although traditional readability measures have proved efficient and relevant in measuring text difficulty, they are limited in considering the multiple aspects of texts (McNamara et al., 2012; Zakaluk & Samuels, 1988). For example, shortening sentences by decreasing the number of propositional phrases, which makes a text easier to understand according to readability formulas, may actually hinder comprehension because there occur more gaps in understanding the meaning of the text. In addition, changing to more familiar words or deleting rare words in the text would risk distorting semantic content and losing literary tastes (Zakaluk & Samuels, 1988).

Moreover, the major drawback of readability measures is that they ignore the discourse features of texts. As noted above in this section, many factors including text content, text organization, or text genre would also influence reading comprehension. Particularly, the way how a text is written

determines considerably how challenging a text is (Alderson, 2000; Armbruster & Anderson, 1985). The texts in which ideas are well-connected tend to be easier to understand than otherwise. Accordingly, as one of the alternatives to unidimensional readability measures, cohesion as ‘the interrelationships of ideas’ (Zakaluk & Samuels, 1988) will be discussed in the next section.

2.3 Cohesion

Halliday and Hasan (1976) defined cohesion in their landmark work, *Cohesion in English* as a semantic one, which indicates “relations of meaning that exist within the text, and that define it as a text” (p. 4). They added as follows.

Cohesion occurs where the INTERPRETATION of some elements in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it. When this happens, a relation of cohesion is set up, and the two elements, the presupposing and the presupposed, are thereby at least potentially integrated into a text (p. 4).

As cited above, cohesion enables a set of sentences to construct a

semantic unit. In other words, a group of unrelated sentences do not create a text. Cohesion is associated more with non-structural, inter-sentential relationship than with structural, intra-sentential relationship because sentences are internally cohesive due to grammar. Such inter-sentential feature has made cohesion as a crucial element in discourse analysis.

Cohesion generally works in the unit of 'tie' which indicates a pair of the referring item and the item that it refers to (Halliday & Hasan, 1976). The relations of ties are primarily anaphoric in which "the presupposing" points back to previous items as "the presupposed." The anaphoric relation often forms a cohesive chain which is a sequence whereby the presupposing refers back to multiple items earlier in the text. On the other hand, in cataphoric relations, the presupposing item, such as 'this' and 'here,' points forward the presupposed. Taking advantage of the 'tie,' Halliday and Hasan presented the way to analyze a text in terms of cohesive properties.

Cohesion is realized by lexicogrammatical system (Halliday and Hasan, 1976). That is, cohesion is expressed by grammar - reference, substitution, ellipsis, and conjunction - and vocabulary. The distinction between grammatical and lexical cohesion is one of degree, and conjunction lies on the borderline between them. Reference generally functions within noun phrase, and there are three types of references: personal (e.g., 'he', 'they' and 'you'), demonstrative (e.g., 'this', 'here' and 'the'), and comparative (e.g., 'same', 'better', 'so' and

‘otherwise’). While reference indicates the relation between meanings for which the text incidentally has the presupposed, substitution is the relation between words or phrases in the text. In other words, the relations of substitution must be within the text. The three types of substitutions are nominal (e.g., ‘one’, ‘ones’ and ‘same’), verbal (e.g., ‘do’), and clausal (e.g., ‘so’ and ‘not’). Ellipsis is also a form of substitution where the item is omitted as in the instance of “Four other oysters followed them, and yet another four (Halliday and Hasan, 1976, p.148).” Conjunction is of semantic relations like reference, but it does not demand searching for the presupposed. Instead, conjunction specifies a way in which sentences connect, for example, additive, adversative, causal or temporal relations. The three kinds of conjunctive expressions are adverbs (e.g., ‘but’, ‘next’ and ‘therefore’), compound adverbs (e.g., ‘furthermore’ and ‘nevertheless’) and prepositional expressions (e.g., ‘as a result of that’ and ‘in spite of that’). Lastly, lexical cohesion is realized through the reiteration of vocabulary (e.g., ‘synonym’ and ‘superordinate’) or the collocation which indicates the regular co-occurrence of lexical items. In sum, cohesion in English consists of lexicogrammatical meaning (lexical cohesion, substitution and ellipsis), referential meaning (reference), and the semantic connection with the preceding text (conjunction).

Though Halliday and Hasan’s explanation of cohesion has been very useful, it is inadequate to fully account for text connectedness. As Carrell (1983)

pointed out, text processing is an interactive process between the text and the reader. Now, the dominant view is that text connectedness is not established solely by the text itself. Rather, text connectedness is of cognitive nature, which indicates that a reader creates his/her own coherent mental representations. Thus, text cohesion would not always lead to the coherent text representation (Carrell, 1982; Xi, 2010). Coherence lies in the mind of the reader, while cohesion lies in the text. For this reason, coherence can only be measured in indirect ways whereas cohesion can be measured directly (McNamara et al., 2014)

A plethora of research studies (Beck et al., 1984; Beck et al., 1991; Britton & Gülgöz, 1991; McNamara et al., 1996) have shown the effects of cohesion on text comprehension and recall. Researchers revised the texts by adding or deleting cohesive cues like references or connectives. Beck et al. (1984) investigated the effects of revised narrative texts on children's comprehension. To increase the ease of the texts, they added clarifications or elaborations in addition to manipulating references and connectives. They found that readers benefitted from revised texts, with skilled readers showing greater benefits compared to less skilled readers. Beck et al. (1991) extended these findings by making the causal connections explicit. This study also confirmed the positive effects of text revision on comprehension. Moreover, as an attempt to modify texts based on the theory of text processing, Britton and Gulgoz (1991) identified 'coherence breaks,' where inferences are needed to fill in the

gaps. Then, they repaired the breaks by adding argument overlaps, rearranging parts of each sentence for old information to come before new information, and making ambiguous references explicit. They did not add extra information to make up the deficit of the reader's knowledge, unlike Beck et al. (1984) and Beck et al. (1991). Their study also showed that the revision led to a better and efficient comprehension. Kobayashi (2002) examined the effects of text organization on English reading comprehension tests for Japanese college students. She found that the more proficient students' performance improved when they read the tightly-organized texts while the less proficient students showed little improvement. Hence, she argued, well-structured texts would distinguish students with different levels of proficiency. In sum, as the above studies have evidenced, increased cohesion tends to facilitate comprehension whereas low cohesion produces more cohesion breaks which become a cognitive burden for a reader (McNamara et al., 2014).

Moreover, the effect of cohesion on text understanding is known to interact with the reader's background knowledge (Best, Floyd, & Mcnamara, 2008; Graesser, McNamara, & Louwerse, 2003; McNamara et al., 2012; McNamara et al., 1996; McNamara & Kintsch, 1996). When encountering cohesion gaps, a reader generally consults prior text or resort to his/her relevant knowledge available (McNamara et al., 2014). Low-knowledge readers might not infer successfully because they do not have sufficient knowledge to repair

the gaps. However, high-cohesion texts would be beneficial to low-knowledge readers. As for high-knowledge readers, the interaction of cohesion with background knowledge is rather complex. For text-based tasks, such as recalling, high-knowledge readers benefited from high-cohesion texts than from low-cohesion texts. On the other hand, for deeper levels of comprehension tasks, such as inferencing or sorting concepts, high-knowledge readers performed better on low-cohesion texts than on high-cohesion texts (McNamara et al., 1996). The so-called ‘reverse effect’ can be explained in terms of Kintsch’s Construction-Integration model (Kintsch, 1988). In the model, the more inferences a reader makes, the deeper text representation he/she would construct. Likewise, in the study of McNamara and her colleagues (1996), low-cohesion texts led high-knowledge readers to make more inferences, resulting deeper understanding.

2.4 Text Length

Intuitively, one would think that longer texts are more difficult to read than shorter ones because they have a greater amount of information to interpret and memorize. Furthermore, the youth of nowadays, who have grown up with the multimedia, often avoid reading longer texts (Oblinger & Oblinger, 2005). Many standardized reading tests also favor short texts. They often contain

numerous brief passages with a wide range of topics so that the effects of reader's background knowledge would be minimized (Alderson, 2000; Ozuru et al., 2008).

However, many reading researchers argue that better reading assessment should use longer passages because longer texts are more effective to induce various aspects of reading comprehension processes. For example, Johnston (1984) argued that since prior knowledge undoubtedly plays an import role in reading comprehension, to eliminate the effects of prior knowledge by using the passages with various content domains is often difficult and undesirable. Instead, he suggested, the extent of background knowledge on the part of a test taker should be measured and considered in interpreting the individual's test score. Moreover, he highlighted that reading assessment is required to use longer texts because "longer texts allow more structure to be built into them and they have greater ecological validity (p.237)." Similarly, Rawson and Kintsch (2005) contended that short texts do not tap into the full range of reader's ability to comprehend a text, for example, the ability to understand text organization. Van den Broek (2012) also maintained that when reading lengthy texts, readers are involved in different levels of complexity, such as identifying the relations of ideas or integrating the ideas scattered across the text. Moreover, Keenan (2012) explained that longer texts provide more textual clues to compensate for decoding deficits whereas the failure to decode could be catastrophic in shorter

texts. Hence, she argued that mental processes involved in understanding longer passages should be the target of reading assessment rather than the reader's ability to decode.

Apart from the importance of longer passages for reading assessment, the experimental research which investigated the effects of longer texts on reading comprehension or learning has shown mixed results. One of the first studies relating to text length was conducted by Newsom and Gaite (1971). In their study, adults were given a short or a long version of texts and took a multiple choice test immediately after and one week after the reading task. The results showed that the subject's performance was better for a short text in the prolonged test while there was no significant difference between the two versions in the immediate test. Likewise, Rothkopf and Billington (1983) found that in self-paced reading, the subjects better recalled for short passages than for long ones. They concluded that the result was partly because the readers failed to integrate related information in longer passages. In Freedle's review of the studies on TOEFL item difficulty (1997), it was also found that passage length has a negative effect on comprehension when it comes to main idea identification. On the other hand, Commander and Stanwyck (1997) found that readers' comprehension monitoring was less successful for shorter texts compared to longer texts. That is, "illusion of knowing," the conflict between the reader's perceived level of understanding and his/her text

comprehension, is more likely to occur with shorter texts whereas longer texts with increased processing time, meaningful context, and abundant textual information would elicit more accurate comprehension monitoring by the reader.

Researchers explained that readers engage in different strategies depending on text length. Engineer (1977 cited in Alderson, 2000) found that the readers' abilities involved in reading lengthy texts are discourse processing abilities rather than syntactic or lexical knowledge. In the same vein, Surber (1992) investigated the effects of passage length on college students' retention, reading speed, and highlighting patterns. He found that readers spend more time per word and highlighted more densely for shorter texts. However, for recall, the results were mixed. While the history text was better recalled for a short version, the economics text was better recalled for a long version. He concluded that when reading shorter texts, readers utilize different strategies from when reading longer texts.

Many studies have also investigated the effects of text length on reading comprehension for second and foreign language learners. Mehrpour and Riazi (2004) found that Iranian college students showed no significant difference between a shorter and an original, longer version in understanding TOEFL texts, which was measured by multiple-choice items. They asserted that the non-significant effect is possibly due to the high proficiency level of the subjects.

Extending this research, Jalilehvand (2012) examined the effects of text length and picture on Iranian high school students' reading comprehension. Although the subjects performed better on the longer text, there was no statistically significant difference between the original version and the shorter version. That is to say, text length had no significant effect on reading comprehension whereas the presence of a picture had a significant effect on text representation.

Simplification procedure is often used in shortening an original text as in the research by Mehrpour and Riazi (2004) and Jailhevand (2012). Many researchers have examined the effects of text simplification or elaboration on reading comprehension of L2 learners since there has been a divide over the use of authentic texts versus simplified reading texts, particularly for learners at the beginning and intermediate levels (Crossley et al., 2007). Though varied across studies, simplification procedure generally involves decreasing the length of sentences and the number of embedded clauses, and replacing multisyllabic, low-frequency words for higher frequency words. Elaboration, in contrast, includes increasing redundancy, signaling devices, and synonyms of low-frequency words (Oh, 2001).

In some studies, simplified texts enhanced reading comprehension for L2 learners. Brantmeier (2005) found that the addition of analogies to scientific texts did not enhance reading comprehension compared to the non-analogy texts. While both versions had no difference for the sentence completion and multiple

choice tests, the non-analogy version was better for the recall task. As L2 readers focused on decoding words and sentences, Brantmeier argued, analogy only hindered comprehension due to the added linguistic features. However, he also cautioned that the addition of sentences hampered the comprehension of details more than that of main idea. In Keshavarz, Atai, and Ahmadi (2007)'s study, the simplified versions were beneficial to understanding of content-unfamiliar text while they had no significant effect on understanding of content-familiar text. Hence, the authors emphasized that teachers should consider the reader's background knowledge in assessing text difficulty.

However, other studies found that there was no significant effect of simplification on text understanding for L2 learners. Yano et al. (1994) used both simplified and elaborated texts to examine the effects of different versions on Japanese college students' reading comprehension. They found that the simplified versions were no better than the elaborated versions for comprehension although the elaborated texts were more linguistically difficult based on conventional readability measures. They inferred that the complexity and greater length of longer texts were offset by the semantic details they provide, thus benefitting comprehension. Similarly, Young (1999) found that simplified texts were not necessarily more effective than authentic texts for recall. In addition, the subjects in her study showed fewer misunderstanding for the longest text than for the shortest one. Like Yano et al. (1994), she concluded

that the cognitive constraints from longer texts might have been neutralized by the redundancy and repetition in the texts. Oh (2001), extending the study of Yano et al., also found that elaborated texts were at least equally successful for reading comprehension. Furthermore, only elaborated input improved the subjects' performance on inference task while both elaborated and simplified version positively influenced the performance of general or specific comprehension test items. Hence, she concluded, elaborated texts would be beneficial for L2 reading because of their rich, native linguistic qualities.

In addition, the recent trend supports the use of authentic texts rather than simplified texts due to the pedagogical trend toward communicative language teaching. The authenticity argument favors original longer texts since people are likely to encounter the texts of greater length in academic and professional settings (Alderson, 2000; Crossley et al., 2007; Ozuru et al., 2008).

Furthermore, the authentic texts replete with elaboration and redundancy are generally considered more cohesive than simplified versions, which help a reader build a coherent text representation. Crossly et al., (2007), who analyzed the various linguistic features in simplified and authentic texts in a computational manner, found that authentic texts contain more connectives, causal verbs and particles compared to simplified texts. In addition, the texts with greater degree of repetition and elaboration are likely to be long. That is, the texts with more cohesive devices, such as conjunctions or similar words,

tend to be long. Discourse coherence, as Haswell (1988) pointed out, is often realized as redundancy, and there is an inter-connection of cohesion and text length.

2.5 Coh-Metrix

Since text comprehension involves multiple facets, such as words, syntax, the textbase and situation model, genre and rhetorical structure, and pragmatic communication level, there has been criticism over traditional readability measures for their unidimensional perspective (McNamara et al., 2014). Thus, the alternatives to analyze the various characteristics of texts were developed, but they often have some restrictions. For example, Halliday and Hasan's analysis of texts are subject to an analyst's judgement and prior experiential knowledge (Zakaluk & Samuels, 1988). In addition, such analysis which greatly depends on human judgment is not easy to apply at large scales.

Accordingly, researchers have tried to devise a way to measure the features of texts objectively and computationally. Such endeavor led to a variety of text analyzing tools which assess beyond word and sentence difficulty. However, the tools which are scattered over separate sources often measure the limited features of texts. For this reason, a more comprehensive platform to analyze the multiple aspects of texts at one stop was needed. In response to the

need to encompass the various computational measures of text analysis, Coh-Metrix has been developed. Coh-Metrix is the product of interdisciplinary research including computational linguistics, corpus linguistics, discourse processes, cognitive science and psychology (McNamara et al., 2014; McNamara et al., 2010).

Coh-Metrix is loaded with the various tools to analyze texts at multiple aspects of words, sentences, and discourse, for example, *WordNet*, the *MRC Psycholinguistic Database*, the *CELEX Lexical Database*, the *type-token ratio*, *vocd*, Measure of Textual Lexical Diversity for lexical diversity, the Charniak parsers, and *Latent Semantic Analysis*.

The tool consists of several major banks of indices; descriptive indices, referential cohesion, latent semantic analysis, lexical diversity, connectives, situation model, syntactic complexity, word information and so on. Under each bank, there are several indices. For instance, descriptive indices include the basic information about the text, such as the number of paragraphs and the number of words. The category of connectives has indices for different types of connectives: causal, logical, contrastive, temporal, and additive. The most recent version (3.0) of Coh-Metrix provides 106 indices in total, including the standard deviations for some measures and traditional readability measures like Flesch Ease and Flesch-Kincaid Grade Level.

While dealing with multiple aspects of texts, the developers of Coh-

Metrix have primarily focused on text cohesion. They have argued that cohesion should be essentially considered in measuring text difficulty since text cohesion is an important factor in predicting how coherent a reader's text representation will be (McNamara et al., 2010; McNamara et al., 2014). In addition, from the practical perspective, there was a growing need for an alternative to conventional readability formulas because the efforts to revise school texts based on readability measures often end up with less coherent texts (Graesser et al., 2004). Furthermore, the developers have assumed that cohesion can be directly measured and quantified. Thus, Coh-Metrix has been designed to provide a variety of measures to assess cohesion, for instance, referential cohesion, LSA cohesion, connectives, and causal cohesion.

Thanks to a group of cohesion measures, Coh-Metrix has been effectively used to examine text cohesion. First of all, McNamara and her colleagues (2010) investigated the validity of Coh-Metrix as a tool to measure text cohesion, using the texts from some published psychology studies. They confirmed that Coh-Metrix significantly distinguishes between high- and low-cohesion texts. In addition, they found that Flesch-Kincaid Grade Level assesses low-cohesion texts as being less difficult than high-cohesion texts, which indicates that the readability formula is limited in assessing the discourse levels of text difficulty. Furthermore, Duran, Bellissens, et al. (2007) sorted science texts into easy or difficult, based on Coh-Metrix indices of referential cohesion

and vocabulary accessibility. The college students showed faster reading time and better recall for 'easy' texts than for 'difficult' texts. Duran, McCarthy, Graesser, and McNamara (2007) also examined the correlation of expert rating and Coh-Metrix indices on temporal cohesion of texts, which was found to be statistically significant. Moreover, the authors showed that Coh-Metrix indices relating to temporal cohesion successfully distinguished between different genres of science, history, and narrative texts.

In addition, Coh-Metrix has been utilized in a wide range of studies to investigate cohesion as well as other linguistic features for different types of texts: narrative and expository texts (McNamara et al., 2012); Korean middle school English texts (Jeon, 2011); changes in scientific articles over time (Bruss, Albers, & McNamera, 2004); written and spoken registers (Louwerse, McCarthy, McNamara, & Graesser, 2004); human tutorial dialogues (Jeon & Azevedo, 2007); simplified and authentic texts (Crossley et al., 2007); Journal abstracts of Japanese, American and British scientists (McCarthy et al., 2007). In sum, Coh-Metrix has successfully proved itself as a theoretically grounded tool to measure the multiple features of texts.

As discussed so far, reading comprehension is a complex process in which various text features and reader variables interact. Nevertheless, the traditional readability formulas measure text difficulty without considering the

multidimensional features of reading comprehension processes. Hence, discourse researchers offered text cohesion as one of the alternatives to measure difficulty of texts at a deeper level. Despite the importance of cohesion for text representation, cohesion in the reading passages on English proficiency tests has not been a major concern. Thus, this study intends to investigate the cohesion of the reading passages on standardized English tests, specifically comparing the tests with varying passage lengths.

Chapter 3.

METHOD

This chapter describes the method employed in the present study. Section 3.1 discusses the corpus. Section 3.2 provides the details on the Coh-Metrix indices selected for this study and the analysis procedure. Finally, statistical analysis is described in Section 3.3.

3.1 Corpus

In this section, the selection and cleaning process of the corpus are described. The corpus for this study comprises the texts gathered from the four different kinds of standardized English proficiency tests for non-native speakers: “College Scholastic Ability Test (CSAT),” “Test of English Proficiency developed by Seoul National University (TEPS),” “Jitsuyo Eigo Gino Kentei, Test in Practical English Proficiency (EIKEN),” and “Test of English as a Foreign Language (TOEFL).” CSAT and TEPS are developed by Korean organizations and mainly administered in Korea. Many Koreans also take TOEFL tests produced by American institute. EIKEN, taken primarily by

Japanese students, was also selected for comparison.

First, CSAT is a Korean college entrance exam which is administered by the national organization, “Korea Institute for Curriculum and Evaluation.” The English section has 45 questions in total, 17 for listening comprehension and 28 for reading comprehension. The section has 25 reading passages whose average word count is 148. With the exception of the last two passages, every passage has only one corresponding question. While the section includes the various types of texts including a letter, advertisement, or narrative, most passages are expository, for example, science and social sciences. Some passages provide Korean translation of difficult words. The time allowed for reading comprehension is about 45 minutes after listening comprehension test finishes.

As the English section on CSAT has changed from the two separate tests depending on English proficiency levels to one unified test starting from 2014, those tests since the change were selected for this study. The CSAT corpus included one official test which was administered in November, 2014 and the four official mock tests which were administered two times in 2014 and 2015, respectively. Since “Korea Institute for Curriculum and Evaluation” releases the previously administered tests on its Internet website, the electronic files of each test were downloaded for this study. Afterwards, the texts were cleaned manually by the researcher because some passages had blanks, underlined phrases, or numbers. In sum, the CSAT corpus has 125 texts in total culled from the five test

sets as shown in Appendix 1.

Second, TEPS was developed by Seoul National University Language Education Institute to measure English proficiency for non-native speakers. The test has the four sections of listening comprehension, grammar, vocabulary, and reading comprehension. The reading section has 40 questions with each question having its own reading passage. The average word count of the passages is 87, and the testing time lasts for 40 minutes. The institute explains that requiring test-takers to read many passages in a relatively short time is necessary to assess their internalized language ability (<http://www.teps.or.kr>). Most passages in the reading section are expository writings except for a few cases of letters and narratives.

The TEPS corpus for this study was selected from the collection of the recently administered tests, which was distributed through private publishers. This study consulted the three books which were published by Hackers Champstudy in 2014. As each reading section on TEPS has 40 passages, which is very large compared to 25 for CSAT, 5 for EIKEN, and 3 for TOEFL, the passages for this corpus were selected in a manner of systematic random sampling. Systematic random sampling is effective to assure that the items are evenly distributed (Ahmed, 2009). For this study, the first three reading tests for each book were selected. Then, starting from the first passage in the first test set, every third passage was chosen. By this procedure, 13 or 14 passages were

selected from each test set. As the tests were available only in the printed versions, they were scanned and transferred to electronic text files. The passages with blanks or jumbled order of sentences were cleaned manually by the researcher. Thus, the corpus of TEPS in this study includes 120 passages randomly selected from nine separate reading sections as indicated in Appendix 2.

Third, EIKEN, which is produced by “EIKEN foundation of Japan,” is the most widely used English tests in Japan. According to the official website (<http://www.eiken.or.jp>), the test is also administered in 45 other countries. The test is used to apply for post-secondary education or get academic credits and qualifications for studying overseas. The test is administered three times a year, summer, fall, and winter. EIKEN has seven different grades with pass-or-fail bases: 1 and Pre-1 for university students; 2 and Pre-2 for high school students; 3, 4, and 5 for junior high school students. The tests for each grade have their own unique set of test items. Higher grade tests assess the ability to use English in educational, social, and professional settings while lower grade tests measure the basic knowledge of English. EIKEN has also two stages, one for listening, reading, and writing, and the other for speaking. The former is a traditional paper-and-pencil test while the latter is a face-to-face interview.

For the corpus of this study, 15 EIKEN tests which were administered from 2013 to 2015 have been selected, specifically for grade 1, Pre-1, and 2. As

CSAT, TEPS, and TOEFL are usually aimed at those who finished high school education or more, the EIKEN corpus excluded the grades below Grade 2. Grade 2 is for high school graduates while Grade Pre-2 is for second-year high school students. Every EIKEN test has the reading section with five reading passages, which is preceded by vocabulary section. Each Grade 1 and Grade Pre-1 test has 16 questions while Grade 2 has 20 questions. Unlike CSAT and TEPS, each passage on EIKEN has its own title. The average number of words per passage on EIKEN is 400. The tests on the official website were downloaded and transferred to electronic text files. Then, the passages with blanks were cleaned manually by the researcher. The total number of texts for EIKEN corpus is 75, five from each 15 tests, as shown in Appendix 3.

Lastly, TOEFL, which is produced and administered by “Educational Testing Service (ETS),” was initially developed for non-native speakers who wish to enroll in American universities, but now is also recognized by 9,000 colleges and institutions in more than 130 countries (<http://www.ets.org/toefl>). The Internet-based TOEFL test comprises the four sections of reading, listening, speaking, and writing. The reading section has 36 to 56 questions with the time limit of 60 to 80 minutes, and some tests would include dummy questions for ETS to verify new test items. The passages are academic texts with the average number of words, 695. Similar to EIKEN, each passage on TOEFL has its own title.

As ETS releases only a small number of official tests, the availability of TOEFL reading passages for this study is very limited compared to other types of tests. That is, the total of 53 passages for TOEFL were drawn for this study: eight from ETS Internet website; 15 from the book of *The Official Guide to the TOEFL Test* (Educational Testing Service, 2014a); 15 from the book of *Official TOEFL iBT Tests* (Educational Testing Service, 2014b); 15 from official TOEFL online practice tests (<http://etest.chosun.com>). Of these passages, five from ETS website and six from the Official Guide were in the form of discrete single passages, not belonging to any complete test set. The online practice tests can be purchased through a Korean ETS preferred vendor, Digital Chosun. The printed versions of tests were scanned and transferred to electronic text files. The free online versions of tests on the ETS website were screen-captured and transferred to text files. The online practice tests on the Digital Chosun website were typed manually by the researcher. The summary of TOEFL corpus in this study is described in Appendix 4.

In sum, a total of 373 passages from the recent official test sets for CSAT, TEPS, EIKEN, and TOEFL were selected for this study.

3.2 Tool

In this section, the tool and the procedure to assess cohesion of the passages on CSAT, TEPS, EIKEN, and TOEFL are described.

3.2.1 Coh-Metrix Indices

As discussed in detail in the Literature Review, Coh-Metrix is the product of interdisciplinary efforts to analyze texts automatically. The tool is a prominent member of applied natural language processing which primarily focuses on textual analysis to deal with language-related issues (McNamara et al., 2014). While there were a range of tools already available to analyze texts, for example, Linguistic Inquiry and Word Count, Concordancers, *Latent Semantic Analysis* (LSA), and *WordNet*, Coh-Metrix is a repository of various text analyzing tools. It provides the theoretically grounded measures for numerous features of texts at word, sentence and discourse levels. That is, the tool covers from sophisticated indices like LSA and syntactic parsers to traditional measures like word count or readability formulas. Coh-Metrix has been successfully validated for measuring what it is supposed to measure (Crossley, Salsbury, & McNamara, 2010; McNamara et al., 2010). Moreover, the tool has been used in a variety of research studies to evaluate the features of different text types (Crossley et al., 2007; Duran, McCarthy, et al., 2007; E. Lightman, P. McCarthy, D. Dufty, & D. McNamara, 2007; E. J. Lightman, P. M. McCarthy, D. F. Dufty, & D. S. McNamara, 2007; McNamara et al., 2012). In sum, Coh-Metrix provides the numerous linguistic measures to successfully assess and differentiate a variety of texts at one stop.

As the purpose of this study is to assess the cohesion of the reading passages on CSAT, TEPS, EIKEN, and TOEFL, the indices that measure cohesion have been selected from a wide range of indices of Coh-Metrix 3.0. The previous studies (Duran, Bellissens, et al., 2007; Graesser et al., 2011; Graesser et al., 2004; McNamara et al., 2010) were consulted in choosing the specific variables for this study. For example, McNamara et al. (2010) listed co-reference, LSA, connectives, and causality as a subset of cohesion-related indices.

To discuss the cohesion indices in detail, first, referential cohesion, also called co-reference, computes overlap between words. The referential indices of Coh-Metrix measure the overlap at two levels, between adjacent sentences (local) and between all of the sentences (global). The indices also vary along the explicitness of overlap. The most stringent overlap, Noun Overlap, measures the exact match between nouns (e.g., “table”/ “table”). It does not consider deviation in morphological forms by excluding words which are only different in plurality from each other. Noun Overlap is one of the most discriminant measures between high- and low- cohesion (McNamara et al., 2010). On the other hand, argument Overlap includes the overlapping words differing in plurality (e.g., “cell”/ “cells”). It also occurs in the matching pronouns (e.g., “he”/ “he”), though it does not determine the referents for the pronoun. The term, “argument,” which indicates nouns or pronouns, is from the

contrast with the term, “predicates,” which refer to verbs or adjectives (McNamara et al., 2014). Stem Overlap, more lenient than noun or argument overlap, measures the overlap between a noun and a content word (e.g., nouns, verbs, adjectives and adverbs) which shares a common lemma (e.g., “tree”/ “treed,” “mouse”/ “mousey,” and “price”/ “priced”). While the three types of referential overlaps have binary values, overlap or non-overlap, Content Word Overlap considers the proportion of the overlapping content words. This study only utilizes the overlaps of noun, stem, and argument which were validated by McNamara and her colleagues (2010) to effectively detect the difference between high- and low-cohesion texts.

Second, latent semantic analysis (LSA) computes conceptual similarity between sentences (local), between all pairs of sentences in a text (global), and between consecutive paragraphs. LSA is a mathematical and statistical technique which originated from information retrieval research (Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988; Foltz, Kintsch, & Landauer, 1998; Louwerse, Landauer, McNamara, Dennis, & Kintsch, 2007). The model takes advantage of the human mechanism which can learn language from the available evidence, a training corpus. Although there are several options of knowledge “space” to choose for users, the Touchstone Applied Science Associates (TASA) corpus is the default corpus for Coh-Metrix. The TASA corpus covers over 12 million words from 30,000 documents of a wide range of topics. It has been found that

LSA highly correlates with human performance of standard vocabulary and subject-matter tests, and it also grades essays as well as human rater (Landauer, McNamara, Dennis, & Kintsch, 2013). In LSA, a word does not have meaning on its own, but the meaning is determined by its relations with other words. For example, “hammer” is semantically related to “nail,” “saw,” and “build,” though they are neither synonyms nor hypernyms (McNamara et al., 2014). LSA indices are not binary, but indicate a degree of semantic similarity ranging from 0 to 1. In addition, Coh-Metrix provides LSA measure of old versus new information in a text. Given information is recoverable from information in a text. By computing the LSA vector of a new sentence compared to that of the preceding text, LSA Given/New index provides the additional measure of cohesion. On the whole, LSA measures, compared to referential cohesion measures, have been found to be effective to detect implicit similarity (McNamara et al., 2010). In this study, LSA Local Overlap, LSA Global Overlap, and LSA Given/New are selected for analysis. The overlap between consecutive paragraphs is excluded since most passages on CSAT and TEPS are of a single paragraph.

Third, connectives contribute to cohesion by clarifying the relations between ideas in a text. Coh-Metrix computes the incidence of all connectives (occurrence per 1,000 words). The tool also measures an incidence score for different types of connectives: causal (e.g., “because” and “so”), logical (e.g.,

“and” and “or”), adversative/contrastive (e.g., “although” and “whereas”), temporal (e.g., “first” and “until”), and additive (e.g., “and” and “moreover”). This study selected Incidence of All Connectives to cover various types of connectives.

Lastly, Coh-Metrix is concerned with a reader’s situation model by providing the ratio of causal particles to causal verbs and the ratio of intentional particles to intentional verbs. The particles including connectives, transitional phrases, adverbs, or other signaling devices are needed when there is a break in cohesion. Causal verbs describe the events or actions which caused something to happen. When there are proportionally more particles that connect actions and events in the text, the text is considered more cohesive. On the other hand, when there are many actions or events without appropriate particles in the text, the text is less cohesive. Although causal relations are often limited to science texts with causal mechanisms or narratives with action plot, this study includes Ratio of Causal Particles to Causal Verbs as one of the analyzing indices since many passages are science-related texts on CSAT, TEPS, EIKEN, and TOEFL.

Besides, this study includes readability indices which measure text difficulty in a more common, traditional way. Coh-Metrix provides Flesch Reading Ease and Flesch-Kincaid Grade Level, both of which are based on word and sentence length. Hence, as the words and sentences increase in length, the score of Flesch Reading Ease decreases and the grade level of Flesch-Kincaid

increases (McNamara et al., 2014).

To sum up, this study selected 11 variables to measure cohesion on CSAT, TEPS, EIKEN and TOEFL as shown in Table 3.1.

Table 3.1
Cohesion Indices for Analysis

Category of Measure	Index	
Referential Cohesion	Noun Overlap	Local/ Global
	Argument Overlap	Local/ Global
	Stem Overlap	Local/ Global
Latent Semantic Analysis	LSA Overlap	Local/ Global
	LSA Given/New	
Connectives	Incidence of All Connectives	
Situation Model	Ratio of Causal Particles to Causal Verbs	

3.2.2 Procedure

Coh-Metrix 3.0 is publicly available on the website, <http://www.cohmetrix.com>. Clicking on the “WEB TOOL” button leads to a login page and next, the screen for text input. On the screen are several types of headers, such as “title,” “genre,” “source,” “job code,” and “LSA space” for the identification of a text. The default LSA space is “college level,” and “Job Code” has to be filled in. Each passage to be analyzed for this study is entered in a copy-and-paste manner from a text file. After putting in necessary information, the “Submit” button on the bottom is pressed. Then, Coh-Metrix provides the

measures of 106 indices on the right side to the text box, with the index labels, numerical values, and descriptions for each index. Next, the results can be downloaded in the form of “Excel” file by clicking on “Save Data” button. The tool is occasionally unstable for system error or any uncertain reasons, and it takes more time to process a longer text than a shorter one.

3.3 Data Analysis

To compare cohesion between different kinds of English proficiency tests, namely, CSAT, TEPS, EIEKEN, and TOEFL, statistical analysis was conducted. Taking test types as the between-group factor and each of the 11 Coh-Metrix cohesion indices as the dependent variables, one-way Analysis of Variance (ANOVA) was conducted using IBM SPSS ver. 22.0. Since the equal variances across the four groups were not assumed for all of the indices, the robust test of Welch ANOVA was performed for every index, and the effect size (ω^2) was also estimated. When there was a significant difference between the tests for a cohesion index, post-hoc analysis of Dunnett T3 was additionally conducted. To measure the strength of difference between each pair of tests, the effect size (Cohen's d) was calculated.

Chapter 4.

RESULTS AND DISCUSSION

This chapter presents the results of the present study and discusses the findings. Section 4.1 reports the results of statistical analysis for 11 cohesion indices. The results are discussed in detail in Section 4.2.

4.1. Results

A one-way analysis of variance (ANOVA) was used to examine the cohesion of the reading passages on English proficiency tests in this study. The independent variable indicated the different tests with four groups: CSAT, TEPS, EIKEN, and TOEFL. The dependent variable was each of 11 cohesion indices: Noun Overlap (Local, Global), Argument Overlap (Local, Global), Stem Overlap (Local, Global), LSA Overlap (Local, Global, Given/New), Incidence of All Connectives, and Ratio of Causal Particles to Causal Verbs.

Since the *Levene's F* test revealed that the homogeneity of variance assumption was not met for all of the dependent variables in this study, the *Welch's F* test was performed, with the level of significance at 0.05. The *Welch's*

F test is known to be rigorous in the case of unequal sample sizes or unequal variances. Since the *Welch's F* test was used, the effect size for each dependent variable was estimated in terms of omega squared (ω^2) formula. In addition, when the result of ANOVA was significant, post hoc analysis was conducted. Because the assumption of homogeneity of variance was violated in this study, Dunnett T3 post hoc procedure was performed using an *a priori* alpha level of 0.05. Moreover, Cohen's *d*, was calculated for each pair to estimate the difference.

Before presenting the results for the cohesion indices, Table 4.1 shows the means and standard deviations for the reading passages with respect to Flesch Reading Ease and Flesch-Kincaid Grade level. As indicated below, regarding the word and sentence difficulty, TOEFL and EIKEN Grade 1 were the most difficult among the tests while CSAT was the easiest.

Table 4.1
Means and Standard Deviations for
Flesch Reading Ease and Flesch-Kincaid Grade Level

		Flesch Reading Ease	Flesch-Kincaid Grade Level
CSAT (n=125)	<i>M (SD)</i>	62.45 (15.50)	7.14 (3.07)
TEPS (n=120)	<i>M (SD)</i>	48.36 (11.91)	10.90 (2.27)
Grade 1 (n=25)	<i>M (SD)</i>	38.62 (7.06)	12.31 (1.22)
EIKEN Grade P1 (n=25)	<i>M (SD)</i>	49.90 (9.07)	10.32 (1.61)
Grade 2 (n=25)	<i>M (SD)</i>	62.17 (7.41)	8.39 (1.41)
TOEFL (n=53)	<i>M (SD)</i>	42.90 (8.65)	12.40 (1.62)

For the results of cohesion indices, first, there was a significant difference between the four tests with regard to referential cohesion. Table 4.2 summarizes the results of the analysis for six referential cohesion indices. As predicted, TOEFL with the longest passages was the highest for most indices. For TEPS and EIKEN, however, the results were mixed. Furthermore, contrary to the prediction, CSAT was the lowest among the four tests although its text length is longer than TEPS.

Table 4.2
Means, Standard Deviations, and Welch's ANOVA Results
for Referential Cohesion

		CSAT (n=125)	TEPS (n=120)	EIKEN (n=75)	TOEFL (n=53)	<i>p</i>	ω^2
Noun Local	<i>M</i>	0.15	0.35	0.32	0.50	.000	0.29
	<i>SD</i>	(0.22)	(0.33)	(0.13)	(0.14)		
Noun Global	<i>M</i>	0.15	0.38	0.27	0.36	.000	0.19
	<i>SD</i>	(0.19)	(0.27)	(0.10)	(0.13)		
Argument Local	<i>M</i>	0.27	0.49	0.41	0.58	.000	0.22
	<i>SD</i>	(0.27)	(0.33)	(0.14)	(0.14)		
Argument Global	<i>M</i>	0.25	0.52	0.35	0.45	.000	0.18
	<i>SD</i>	(0.24)	(0.28)	(0.11)	(0.13)		
Stem Local	<i>M</i>	0.22	0.46	0.42	0.62	.000	0.32
	<i>SD</i>	(0.26)	(0.34)	(0.14)	(0.14)		
Stem Global	<i>M</i>	0.22	0.50	0.36	0.49	.000	0.24
	<i>SD</i>	(0.23)	(0.29)	(0.11)	(0.13)		

Regarding Noun Local Overlaps, the one-way ANOVA revealed a statistically significant main effect, *Welch's* $F(3, 185.30) = 51.883, p < .001$. The

omega squared ($\omega^2 = 0.29$) indicated that approximately 29 % of the total variation in average value on Noun Local Overlap is attributable to the differences between the four types of tests. Next, for Noun Global Overlap, there was also a statistically significant difference, *Welch's* $F(3, 179.737) = 30.169, p < .001$.

In addition, post-hoc comparisons were conducted to determine which pairs of the four tests differed significantly regarding Noun Overlaps as shown in Table 4.3. The Dunnett T3 analysis for Noun Local Overlap indicates that there was a significant difference between all pairs of tests except between TEPS and EIKEN. The difference is the largest between CSAT and TOEFL, with an effect size of 1.90. For Noun Global Overlap, all of the tests are significantly different from each other except between TEPS and TOEFL. The difference is the largest between CSAT and TOEFL, with an effect size of -.1.30.

Table 4.3
Dunnett T3 Post hoc Results for Noun Overlap

	CSAT - TEPS	CSAT - EIKEN	CSAT - TOEFL	TEPS - EIKEN	TEPS - TOEFL	EIKEN - TOEFL
Noun Local	-.19 (-.71)*	-.17 (-.94)*	-.35 (-1.90)*	.02 (.12)	-.15 (-.59)*	-.17 (-1.33)*
Noun Global	-.23 (-.99)*	-.11 (-.79)*	-.21 (-1.30)*	.11 (.54)*	.02 (.09)	-.09 (-.78)*

Notes. (1) Mean Differences ($\bar{X}_i - \bar{Y}_j$); (2) The effect size, Cohen's d is indicated in parentheses. * $p < .05$.

With regard to Argument Overlap, the one-way ANOVA of the measure of Argument Local Overlap revealed a statistically significant main effect, *Welch's* $F(3, 189.894) = 36.129, p < .001, \omega^2 = 0.22$. There was also a significant difference for Argument Global Overlap, *Welch's* $F(3, 184.788) = 28.973, p < .001, \omega^2 = 0.18$.

According to the post hoc comparisons for Argument Overlap as indicated in Table 4.4, all of the tests are significantly different from each other for Argument Local Overlap, except the pair of TEPS and EIKEN and the pair of TEPS and TOEFL. The effect size is very large between CSAT and TOEFL (-1.44) and between EIKEN and TOEFL (-1.21). For Argument Global Overlap, all of the tests are significantly different from each other except between TEPS and TOEFL. The difference is large for the pair of CSAT and TEPS and the pair of CSAT and TOEFL, with an effect size of -1.03 and -1.04, respectively.

Table 4.4
Dunnett T3 Post hoc Results for Argument Overlap

	CSAT - TEPS	CSAT - EIKEN	CSAT - TOEFL	TEPS - EIKEN	TEPS - TOEFL	EIKEN - TOEFL
Argument Local	-.22 (-.73)*	-.14 (-.65)*	-.31 (-1.44)*	.09 (.32)	-.09 (-.36)	-.18 (-1.21)*
Argument Global	-.27 (-1.03)*	-.09 (-.53)*	-.19 (-1.04)*	.17 (0.80)*	.07 (.32)	-.10 (-.83)*

Notes. (1) Mean Differences ($\bar{X}_j - \bar{Y}_j$); (2) The effect size, Cohen's d is indicated in parentheses. * $p < .05$.

For Stem Overlap, the one-way ANOVA of the measure of Stem Local Overlap revealed a statistically significant main effect, *Welch's* $F(3, 187.895) = 59.076$, $p < .001$, $\omega^2 = 0.32$. Next, for Stem Global Overlap, there was also a significant difference, *Welch's* $F(3, 183.731) = 39.726$, $p < .001$, $\omega^2 = 0.24$.

The post hoc analysis, as shown in Table 4.5, indicates that for Stem Local Overlap, there was a significant difference between all pairs of tests except the pair of TEPS and EIKEN. The effect size is very large between CSAT and TOEFL (-1.92), and EIKEN and TOEFL (-1.43). According to the post hoc analysis of Stem Global Overlap, all of the tests are significantly different from each other except between TEPS and TOEFL. The effect size is the largest between CSAT and TOEFL (-1.45).

Table 4.5
Dunnet T3 Post-hoc Results for Stem Overlap

	CSAT - TEPS	CSAT - EIKEN	CSAT - TOEFL	TEPS - EIKEN	TEPS - TOEFL	EIKEN - TOEFL
Stem Local	-.25 (-.79)*	-.21 (-.96)*	-.40 (-1.92)*	.04 (0.15)	-.16 (-.62)*	-.20 (-1.43)*
Stem Global	-.29 (-1.07)*	-.15 (-.78)*	-.27 (-1.45)*	.14 (.64)*	.02 (.04)	-.12 (-1.08)*

Notes. (1) Mean Differences ($\bar{X}_i - \bar{Y}_j$); (2) The effect size, Cohen's d is indicated in parentheses. * $p < .05$.

Regarding Latent Semantic Analysis, Table 4.6 summarizes the results

of analysis for three LSA indices. The results support the prediction, with TOEFL having the highest overall means. However, for TEPS and EIKEN, the results are mixed. Similar to the results for referential overlaps, CAST is lowest among the four tests.

Table 4.6
Means, Standard Deviations, and Welch's ANOVA Results
for Latent Semantic Analysis

		CSAT (n=125)	TEPS (n=120)	EIKEN (n=75)	TOEFL (n=53)	<i>p</i>	ω^2
LSA	<i>M</i>	0.15	0.23	0.20	0.28	.000	0.23
Local	<i>SD</i>	(0.11)	(0.11)	(0.05)	(0.06)		
LSA	<i>M</i>	0.13	0.24	0.18	0.26	.000	0.25
Global	<i>SD</i>	(0.11)	(0.11)	(0.05)	(0.06)		
LSA	<i>M</i>	0.27	0.22	0.31	0.36	.000	0.57
Given/New	<i>SD</i>	(0.48)	(0.49)	(0.34)	(0.33)		

The one-way ANOVA of the measure of LSA Local revealed a statistically significant main effect, *Welch's* $F(3, 185.983) = 37.466$, $p < .001$. The omega squared ($\omega^2 = 0.23$) indicated that approximately 23 % of the total variation in average value on LSA local is accounted for by the differences between the four types of tests. For LSA Global, there was a significant difference, *Welch's* $F(3, 184.816) = 42.980$, $p < .001$, $\omega^2 = 0.25$. Lastly, for LSA Given/New, all of the tests were significantly different from each other,

Welch's $F(3, 179.746) = 168.571, p < .001, \omega^2 = 0.57$. The discriminating power of LSA Given/New for the four tests is the greatest among the cohesion indices in this study.

Additionally, Dunnett T3 post hoc procedures were conducted for LSA measures, as given in Table 4.7. The post hoc analysis indicates that for LSA Local Overlap, all of the tests are significantly different from each other except the pair of TEPS and EIKEN. The effect size is very large between CSAT and TOEFL (-1.47), and between EIKEN and TOEFL(-1.45). For LSA Global Overlap, there was a significant difference between the tests except between TEPS and TOEFL. The effect size is very large between CSAT and TOEFL (-1.47), and between EIKEN and TOEFL (-1.45). In addition, the post hoc analysis reveals that LSA Given/New is the only index which differentiates every test from each other. The effect size is very large for all the pairs.

Table 4.7
Dunnett T3 Post-hoc Results for Latent Semantic Analysis

	CSAT - TEPS	CSAT - EIKEN	CSAT - TOEFL	TEPS - EIKEN	TEPS - TOEFL	EIKEN - TOEFL
LSA Local	-.07 (-.72)*	-.04 (-.59)*	-.13 (-1.47)*	.03 (.35)	.05 (-.56)*	-.08 (-1.45)*
LSA Global	-.12 (-1.00)*	-.06 (-.59)*	-.13 (-1.47)*	.06 (.70)*	-.01 (-.23)	-.07 (-1.45)*
LSA Given/New	.05 (1.00)*	-.04 (-.97)*	-.09 (-2.18)*	.09 (-2.18)*	-.14 (-3.40)*	-.05 (-1.67)*

Notes. (1) Mean Differences ($\bar{X}_i - \bar{Y}_j$); (2) The effect size, Cohen's d is indicated in parentheses. * $p < .05$.

As for Incidence of All Connectives, Table 4.8 summarizes the results of analysis. The one-way ANOVA of the measure of Incidence of All Connectives revealed no statistically significant main effect, *Welch's* $F(3, 190.049) = .473$, $p > .05$, which is contrary to the prediction. Incidence of All Connectives is the only index which does not differentiate between the tests. Accordingly, no further analysis of post-hoc was conducted.

Table 4.8
Means, Standard Deviations, and Welch's ANOVA Results
for Incidence of All Connectives

		CSAT (n=125)	TEPS (n=120)	EIKEN (n=75)	TOEFL (n=53)	p	ω^2
Incidence of All Connectives	<i>M</i>	88.90	89.52	87.15	89.88	.702	N/A
	<i>SD</i>	(26.38)	(31.92)	(13.62)	(13.54)		

Finally, for Ratio of Causal Particles to Causal Verbs, Table 4.9 summarizes the results of analysis. The one-way ANOVA of the measure of LSA Given/New revealed a statistically significant main effect, *Welch's* $F(3, 179.658) = 11.647$, $p < .001$, $\omega^2 = 0.08$. The effect size indicates that approximately merely 8 % of the total variation in average value on LSA Global is accounted for by the differences between the four types of tests.

Table 4.9
Means, Standard Deviations, and Welch's ANOVA Results
for Ratio of Causal Particles to Causal Verbs

		CSAT (n=125)	TEPS (n=120)	EIKEN (n=75)	TOEFL (n=53)	<i>p</i>	ω^2
Ratio of Causal Particles to Causal Verbs	<i>M</i> <i>SD</i>	0.25 (0.32)	0.37 (0.44)	0.47 (0.22)	0.41 (0.24)	.000	0.08

Post hoc analysis, as shown in Table 4.10, indicates that most pairs are not significantly different from each other. The effect size is large only between CSAT and EIKEN.

Table 4.10
Dunnett T3 Post-hoc Results for Ratio of Causal Particles to Causal Verbs

	CSAT - TEPS	CSAT - EIKEN	CSAT - TOEFL	TEPS - EIKEN	TEPS - TOEFL	EIKEN - TOEFL
Ratio of Causal Particles to Casual Verbs	-.13 (-.31)	-.22 (-.80)*	-.17 (-.57)*	-.09 (-.29)	-.04 (-.11)	.05 (.26)

Notes. (1) Mean Differences ($\bar{X}_i - \bar{Y}_j$); (2) The effect size, Cohen's *d* is indicated in parentheses. **p* = .000.

4.2. Discussion

This study examined differences in cohesion between the reading

passages on CSAT, TEPS, EIKEN and TOEFL, specifically with respect to 11 cohesion indices of Coh-Metrix. To sum up the results of statistical analysis in this study, there was a significant difference between the tests for all of the cohesion indices except for Incidence of All Connectives.

For the first research question of whether the reading passages on CSAT, TEPS, EIKEN and TOEFL are different from each other with respect to cohesion, the results strongly support the prediction. For example, TOEFL with the longest passages was the highest by most measures. In addition, CSAT with shorter passages was lower for all of the indices than EIKEN and TOEFL, except for Incidence of All Connectives. However, the results for TEPS and EIKEN varied. Although TEPS has much shorter passages than EIKEN, the former was as cohesive as or more cohesive than the latter. Furthermore, CSAT, though its passages were longer than TEPS, was the lowest for most indices among the four tests. Regarding Ratio of Causal Particles to Causal Verbs, the difference was significant, though minimal. For Incidence of All Connectives, there was no significant difference between the tests.

In addition, for most indices, CSAT and TEPS showed greater standard deviations than EIKEN and TOEFL, indicating that the cohesion levels on CSAT and TEPS are very inconsistent across the passages. One of the possible explanations for these results is that very short passages are likely to have upper and lower extreme results. In a brief text, one overlap or two is often enough to

dramatically increase the overall ratio for an index, as illustrated in a TEPS passage below.

- 01 Propaganda is a deliberate attempt to sway an opinion.
- 02 Successful techniques appeal to basic, universal human emotion
and desire, rather than to cold, hard *reason*.
- 03 For example, *propaganda* can cause people to *believe* or do
things they might not have, had they *reasoned* more carefully.
- 04 *Propagandists* are not concerned with good or bad, right or
wrong, but with inciting an action or instilling a common *belief*.
- 05 Everyone is susceptible to the suggestions of *propaganda* on
some level, but those who do not think for themselves are most
strongly influenced by its techniques (TEPS 1-1, no. 31).

In the above passage, there are Stem Local Overlaps between sentences (02-03, 03-04, 04-05) as indicated in italics. That is, only three pairs of overlaps in this text lead to a relatively high ratio of 0.75 for Stem Local Overlap. On the other hand, for much longer passages, having one overlap or two would not make a significant difference to the overall ratio for the index. Likewise, the passages on CSAT and TEPS are more likely to be extremely low in cohesion than those on EIKEN and TOEFL.

Despite the similar wide variations on cohesion, CSAT is lower than TEPS for most cohesion indices. This result is quite contrary to the prediction since the passages on CSAT are longer than those on TEPS. To explain the low cohesion of CSAT passages compared to the rest of the tests, the reader's knowledge level which the authors might have assumed can be considered. In the previous study (McNamara et al., 2012), the cohesion of the science texts was higher than narratives and the social science texts. McNamara and her colleagues supposed that the science texts which often deal with unfamiliar and difficult concepts are likely to be more cohesive in order to facilitate readers' understanding. On the other hand, as the social studies texts in the research were less cohesive, the writers of the social studies texts might have assumed the considerable level of domain knowledge on the part of the readers. To apply such an explanation to the current study, it is worth noting that most passages on CSAT are expository texts which discuss scientific and social phenomena. Accordingly, the low cohesion of the passages on CSAT might indicate that the authors assumed that their readers would have sufficient background knowledge on the topics and thus, have little difficulty with understanding the texts. For the sake of illustration, here is a low cohesion passage from CSAT.

Investigators as a personality type place a high value on science, process, and learning. They excel at research, using logic and the

information gained through their senses to conquer complex problems. Nothing thrills them more than a “big find.” Intellectual, introspective, and exceedingly detail-oriented, investigators are happiest when they’re using their brain power to pursue what they deem as a worthy outcome. They prefer to march to their own beat, and they dislike overly structured environments that necessitate a set response to challenges. Investigators are not interested in leadership, and developing the interpersonal skills necessary to fuel collaboration is a hurdle for many of them. They may feel insecure in their ability to “keep up” in their fields and can react badly when forced to put more important work on hold to complete a task that doesn’t intrigue them (CSAT September, 2016, no.32)

The above passage is of very low cohesion according to many Coh-Metrix indices: Noun Local (0); Argument Local (0.286); Stem Local (0); Noun Global (0.107); Stem Global (0.107); LSA Local (0.145); LSA Global (0.131); LSA Given/New (0.196). Moreover, it uses primarily conceptual words, such as “logic,” “introspective” and “leadership,” which are usually considered more difficult than concrete words. The passage is also likely to be an excerpt because it seems to be a part of the text on a variety of personality types or on the different aspects of investigators. Thus, the limited information in this brief

passage would not be enough for a Korean high school student to clearly understand the text. What is even worse, the test taker is under time-pressure. Put differently, the low- cohesion reading passages on CSAT would be hard for test takers, especially for those with low domain knowledge. Low cohesion in the texts might increase the cognitive demands on the part of a reader particularly when the reading passages are conceptually difficult or unfamiliar to the reader.

For the second research question of which features of cohesion drives differences between the tests, first of all, LSA Given/New distinguished between the tests to the greatest extent among the indices of interest. Every single test is significantly different from each other with respect to LSA Given/New. Most importantly, the results are exactly in accordance with passage length: TOEFL, EIKEN, CSAT and TEPS from highest to lowest. Usually, when a reader encounters a new sentence, he/she processes the sentence based on the prior text information. Thus, LSA Given/New compares the conceptual similarity between the current sentence and the previous text. That is, the index calculates the proportion of given information to the total of given and new information: $\text{Given} / (\text{New} + \text{Given})$. The longer the text, the more old information it is likely to offer. In other words, longer texts provide a reader with greater amount of information to help understanding. Hence, the results for LSA Given/New also confirm that a test taker reading shorter passages does not have plenty of text resources to consult. When reading shorter texts, as Keenan (2012) argued, the effects of the

failure to decode a couple of words or sentences could be catastrophic due to the scarcity of textual information.

However, lower cohesion in some aspects could be offset by higher cohesion in other aspects. In McNamara et al. (2012), low referential cohesion in the narrative texts seemed to be compensated for by high verb cohesion. In other words, as McNamara and her colleagues argued, the authors may negotiate between the features of text difficulty, increasing the difficulty on the one hand and decreasing the difficulty on the other. Similarly, for TEPS in the present study, the low level of LSA Given/New seems to be compensated for by the high level of reference and LSA cohesion. That is, although the reader cannot recover plenty of old information from the text, he/she can still take advantage of referential and LSA overlaps. By contrast, CSAT is low by most cohesion measures, implying a greater comprehension challenge for a reader.

The next biggest difference between the tests was found for Stem Local Overlap and Noun Local Overlap. The results for Noun Local Overlap were similar to the previous studies (McNamara et al., 2010; McNamara et al., 2006). Noun Local Overlap, which counts the exact match of nouns only, was very discriminative between high- and low- cohesion texts. What is interesting in the current study is that Stem Overlap, which is rather lenient, provided slightly greater distinction than Noun Overlap, as shown in the bigger effect sizes.

For both Stem Local Overlap and Noun Local Overlap, TOEFL was

significantly higher than the rest of the tests. These results suggest that a greater portion of adjacent sentences on TOEFL passages is lexically connected together although TOEFL has much more sentences than the rest of the tests. As written in italics in the following TOEFL passage, many pairs of consecutive sentences have Noun or Stem Overlaps. In addition, unlike the passage without a title on CSAT and TEPS, the passage on TOEFL could have an overlap in the first sentence with its title.

Electricity from Wind

Since 1980, the use of *wind* to produce *electricity* has been growing rapidly. In 1994 there were nearly 20,000 *wind* turbines worldwide, most grouped in clusters called *wind* farms that collectively *produced* 3,000 megawatts of *electricity*. Most were in Denmark (which got 3 percent of its *electricity* from *wind turbines*) and California (where 17,000 machines *produced* 1 percent of the state's *electricity*, enough to meet the residential needs of a city as large as San Francisco). In principle, all the power *needs* of the United States could be provided by exploiting the *wind* potential of just three *states*, North Dakota, South Dakota, and Texas.

Large *wind* farms can be built in six months to a year and then easily expanded as needed. With a moderate to fairly high net energy yield, these systems emit no heat-trapping carbon dioxide or other air

pollutants and *need* no water for cooling; manufacturing them produces little water pollution. The land under wind turbines can be used for grazing cattle and other purposes, and leasing land for wind turbines can provide extra income for farmers and ranchers.

Wind power has a significant cost advantage over nuclear power and has become competitive with coal-fired power plants in many places. With new technological advances and mass production, projected *cost* declines should make *wind power* one of the world's cheapest ways to produce electricity. In the long run, *electricity* from large *wind* farms in remote areas might be used to make hydrogen gas from water during periods when there is less than peak demand for *electricity*. The *hydrogen gas* could then be fed into a storage system and used to generate *electricity* when additional or backup power is needed.

Wind power is most economical in areas with steady winds. In areas where the *wind* dies down, backup electricity from a utility company or from an energy storage system becomes necessary. *Backup* power could also be provided by linking *wind* farms with a solar cell, with conventional or pumped-storage hydropower, or with efficient natural-gas-burning turbines. Some drawbacks to *wind farms* include visual pollution and noise, although these can be overcome by improving their design and locating them in isolated areas.

Large *wind farms* might also interfere with the flight patterns of migratory birds in certain *areas*, and they have killed large birds of prey (especially hawks, falcons, and eagles) that prefer to hunt along the same ridge lines that are ideal for *wind* turbines. The *killing* of *birds of prey* by *wind turbines* has pitted environmentalists who champion wildlife protection against environmentalists who promote renewable *wind* energy. Researchers are evaluating how serious this problem is and hope to find ways to eliminate or sharply reduce this problem. Some analysts also contend that the number of birds killed by wind turbines is dwarfed by birds killed by other human-related sources and by the potential loss of entire bird species from possible global warming. Recorded deaths of *birds of prey* and other *birds* in *wind* farms in the United States currently amount to no more than 300 per year. By contrast, in the United States an estimated 97 million *birds* are killed each year when they collide with buildings made of plate glass, 57 million are killed on highways each year; at least 3.8 million *die* annually from pollution and poisoning; and millions of *birds* are electrocuted each year by transmission and distribution lines carrying power produced by nuclear and coal power plants.

The technology is in place for a major expansion of wind *power* worldwide. *Wind power* is a virtually unlimited source of energy at

favorable sites, and even excluding environmentally sensitive areas, the global potential of *wind power* is much higher than the current *world* electricity use. In theory, Argentina, Canada, Chile, China, Russia, and the United Kingdom could use *wind* to meet all of their *energy* needs. *Wind power* experts project that by the middle of the twenty-first century *wind power* could supply more than 10 percent of the world's electricity and 10-25 percent of the electricity used in the United States (www.toefl.org).

In contrast, CSAT was the lowest among the tests for Noun and Stem Local Overlap in addition to the other types of referential cohesion. These results imply that the fewer overlaps of words on CSAT passages might often force a reader to infer in order to repair the cohesion gaps by using background knowledge. For example, in the following CSAT passage, there are few Noun or Stem Local Overlaps between adjacent sentences as indicated in italics.

People make extensive use of searching images. One unexpected context is sorting. Suppose you have a bag of small hardware — screws, nails, and so on — and you decide to organize them into little jars. You dump the stuff out on a table and begin separating the items into coherent groups. It is possible to do this by

randomly picking up individual objects, one by one, identifying each one, and then moving it to the appropriate jar. But what most people do is very different. They quickly pick out a whole series of items of the same type, making a handful of, say, small screws. They put them in the jar and then go back and do the same for a different kind of *item*. So the sorting sequence is nonrandom, producing runs of *items* of a single type. It is a faster, more efficient technique, and much of the increased efficiency is due to the use of searching images (CSAT, June, 2016, no.39).

For TEPS and EIKEN, there was no significant difference between them with regard to the local referential overlaps. Put differently, TEPS is as cohesive as EIKEN although TEPS passages are much shorter than EIKEN passages.

Concerning global referential overlaps, it is interesting that TEPS is often higher for Global Overlaps in both referential and LSA indices. Particularly, the means of TEPS for global referential overlaps were even higher than those of TOEFL, although the difference was not statistically significant. On the contrary, the rest of the tests in the present study and the texts from the previous research (McNamara et al., 2010; McNamara et al., 2006) were higher for local cohesion than for global cohesion. These results may have been because the passages on TEPS are very short. Only one or two occurrences of overlaps

between non-adjacent sentences in a very brief text could greatly increase the overall ratio of global overlap. Another explanation might be that the passages on TEPS have more referential or conceptual overlaps across the text compared to the rest of the tests. As illustrated below, every sentence has a lexical overlap with another sentence in the text.

Mammals have two ways to remove excess *salt* from ingested food. Their *kidneys* are effective at removing *sodium*, and *sweat glands* eliminate anything left over through the *skin*. However, *reptiles* such as the marine iguana and sea turtle have their own effective system of removing excess *salt*. Their *kidneys* are inadequate for filtering *sodium* out of their marine-based diets, and *salt* is not removed through the *skin* because they do not have *sweat glands*. To compensate for this, these *reptiles*, through the process of evolution, have developed *salt glands*, which are specifically dedicated to excreting excess *sodium* (TEPS 2-3, no.6).

Furthermore, contrary to the prediction based on passage length, TEPS is as cohesive as or more cohesive than EIKEN for most indices. One plausible explanation for this result is, again, that since TEPS passages are very short, they might have extremely high values for some indices. Another explanation would

be that the authors of EIKEN passages did not take advantage of text length well enough to surpass TEPS in terms of cohesion. The following histogram, Figure 1, shows the distribution pattern of Stem Global Overlap for TEPS and EIKEN. The range of the values for TEPS is wider than EIKEN. In addition, the median value for TEPS is 0.50 while that for EIKEN is 0.35, which suggests that TEPS passages are often higher in cohesion than EIKEN passages.

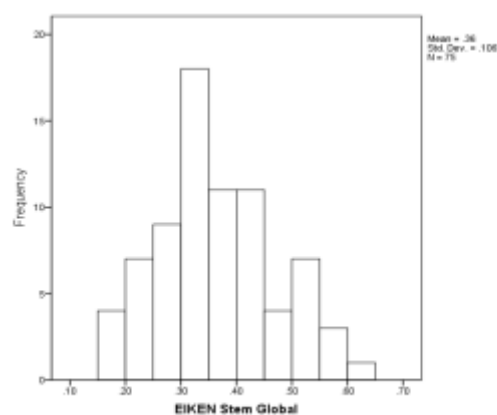
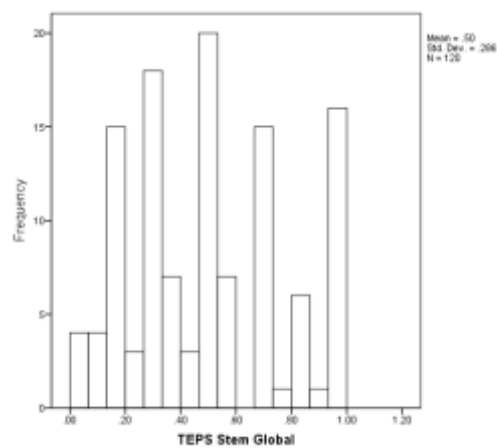


Figure 4.1 Distribution of Stem Global Overlap for TEPS and EIKEN

For both LSA Local and LSA Global overlap indices, they generally followed the patterns found for referential cohesion, with TOEFL being the highest, CSAT being the lowest, and TEPS and EIKEN being in-between. LSA overlaps and referential overlaps are related variables (McNamara et al., 2010), but LSA overlaps are usually effective to detect implicit, conceptual similarity while referential overlaps are more sensitive to explicit linguistic similarity. For this reason, LSA overlaps are often generous in its determination of overlap and thus, less discriminative between high- and low-cohesion texts than referential overlap (McNamara et al., 2010). In this study, however, it was found that LSA overlaps are as much powerful as referential overlaps in distinguishing between high- and low-cohesion texts. For example, the difference between CSAT and TOEFL for both LSA Overlaps was as large as that for referential overlaps, which indicates that CSAT is less cohesive than TOEFL, explicitly and implicitly. These results may also suggest that the passages on CSAT would be more challenging to comprehend when a reader does not have sufficient knowledge to fill in the cohesion gaps.

For Connectives, TEPS and TOEFL were slightly higher than CSAT and EIKEN, though there was no significant difference between the tests. There is no gold standard for the occurrence of connectives, but the connectives were found more often in this study (88.86 on average) than in other studies: 73.26 for high-cohesion texts and 69.29 for low-cohesion texts (McNamara et al., 2006); 67.07

for science texts and 69.10 for history texts (McCarthy, Lightman, Dufty, & McNamara, 2006); 79.1 for the scientific articles (Bruss et al., 2004).

Ratio of Causal Particles to Causal Verbs in this study was low for all the tests compared to the previous studies (McNamara et al., 2006; McNamara et al., 2010). McNamara et al. (2010) pointed out that Causal Particle Ratio is unstable when the text is very short or when the text has very few causal verbs. In the previous research (McNamara et al., 2006), the difference between high- and low-cohesion texts was minimal when there was no explicit causal cohesion manipulation. In the current study, there was a significant difference between the tests, though small. Particularly, CSAT was the lowest while the rest of the tests were not significantly different from each other. The result indicates that the causal relations on the passages on CSAT, if any, are the least explicit among the four tests.

Finally, according to Flesch Reading Ease and Flesch-Kincaid Grade Level, TOEFL was the most difficult whereas CSAT was the easiest among the tests. However, this study also found that TOEFL was more cohesive than CSAT with respect to many cohesion indices. These results suggest that the high cohesion of the TOEFL passages might help reading comprehension whereas the low cohesion of the CSAT passages might hamper text understanding, especially for low-knowledge readers. Hence, this study confirms that the traditional readability measures based only on word and sentence difficulty might not fully

explain text comprehensibility.

In sum, the reading passages on English proficiency tests in this study differed from each other in terms of cohesion. As expected, TOEFL with the longest passages was often the most cohesive of all. However, for the rest of the tests, the results were mixed. EIKEN was as cohesive as or less cohesive than TEPS although its passages are much longer than those of TEPS. In addition, CSAT was the least cohesive among the four tests despite its greater length than TEPS. Thus, while adding cohesion to a text tends to increase text length, longer texts are not always more cohesive than shorter texts.

Furthermore, the cohesion of the reading passages on some tests was very low when compared to other tests. Low-cohesion texts are more likely to force the reader to use his/her background knowledge to fill in the cohesion gaps. In particular, when the topics of the passages are unfamiliar or challenging to the reader as in expository texts, low cohesion might seriously hamper text understanding for low-knowledge readers. Then, the validity of the reading comprehension test might be doubtful because it is uncertain if the poor performance of the test taker is caused by his/her poor reading ability or the deficit of relevant background knowledge. Thus, in developing the passages for reading assessment, it is essential to consider the cohesion of the texts in addition to word or sentence difficulty.

Chapter 5.

CONCLUSION

This chapter is composed of three sections. Section 5.1 summarizes the major findings of the present study. In Section 5.2, the implications for reading assessment are presented. Finally, Section 5.3 reports the limitations of the current study and makes suggestions for further research.

5.1 Major Findings

This study examined the reading passages on English proficiency tests, CSAT, TEPS, EIKEN, and TOEFL with respect to 11 cohesion indices of the automated text analysis tool, Coh-Metrix. Based on the previous research (Crossley et al., 2007; Graesser et al., 2011; Haswell, 1988), this study predicted that cohesion would be associated with passage length for the corpus of this study. In other words, the researcher expected that longer passages would be more cohesive than shorter passages.

Regarding the first research question of whether the passages on CSAT, TEPS, EIKEN, TOEFL are different from each other with respect to cohesion,

there was a significant difference between the tests for all of the indices except Incidence of All Connectives. In line with the expectation, TOEFL with the longest passages was the highest for most indices. However, for the rest of the tests, the results were mixed. The passages on TEPS were as cohesive as or more cohesive than the EIKEN passages, which is contrary to the prediction. Moreover, the CSAT passages were the lowest for most indices although they are longer than the TEPS passages. The EIKEN passages were more cohesive than the CSAT passages as predicted. In addition, it was found that the cohesion of the passages on CSAT and TEPS was very inconsistent compared to the EIKEN and TOEFL passages.

For the second research question of what features of cohesion drives the difference between the tests, all of the cohesion indices selected for this study distinguished between the tests apart from Incidence of All Connectives. Particularly, LSA Given/New was the most discriminative among the indices, which strongly supports the prediction based on passage length. Next, both Noun Local and Stem Local Overlap distinguished between the tests to a great extent. In addition, LSA Overlaps in this study was as much powerful as referential overlaps in differentiating between the tests. For causal cohesion, there was also a significant difference, though minimal.

5.2 Implications

This study intends to draw test developers' attention to the multidimensional features of texts, such as cohesion, so that the text characteristics on standardized English tests could be better understood. Although the effect of cohesion on test-takers performance was not directly investigated in this study, the previous research on cohesion and text understanding (Beck et al., 1984; Beck et al., 1991; Britton & Gülgöz, 1991; McNamara et al., 2014; McNamara et al., 1996) suggests that text comprehensibility might be different from the expectation based on traditional readability measures or the common perception surrounding text length. For example, some easy texts according to word and sentence difficulty might actually be more difficult if they are less cohesive. In addition, longer passages of high-cohesion can be easier to understand than shorter passages of low-cohesion. In particular, low-cohesion texts would be a trouble for low-knowledge readers who do not have sufficient knowledge to bridge the cohesion gaps. Obviously, readers sometimes need to be challenged by difficult texts, but if background knowledge is not the intended construct of reading assessment, the passages must be written cohesively so as not to seriously impede the low-knowledge readers' understanding. Hence, the appropriate selection of reading passages becomes critically important for the accurate assessment of reading ability, especially for high-stakes exams.

In addition, longer texts should be more often used for reading assessment. As implied in the present study and also found in the previous research, longer passages are not always more difficult to understand than shorter passages. Since understanding longer passages usually involves a wider range of comprehension processes compared to shorter passages, they are more effective in assessing the sophisticated aspects of reading comprehension, such as understanding the organization of the texts and integrating the ideas scattered across the text. Moreover, the authenticity argument in communicative language teaching lends support to the use of longer passages in reading assessment (Crossley et al., 2007). The texts which English learners encounter in academic and professional settings are more likely to be longer than those in CSAT and TEPS.

In conclusion, this study underscores that in developing passages for reading assessment, test designers would do well to consider the deeper levels of text, such as cohesion, instead of depending on surface linguistic features only. Furthermore, test developers should take into account of how the background knowledge of a test-taker would interact with the cohesion of the texts. In other words, for the sake of good reading assessment, the text features should be considered in concert with the reader characteristics.

5.3 Limitations and Suggestions

The results of this study which used Coh-Metrix indices should be interpreted with caution since the quantitative analysis are limited in exhaustively detecting the features of text cohesion. Low cohesion texts according to some Coh-Metrix indices can probably be coherent in other aspects, as found in McNamara et al. (2012). Furthermore, the present study only speculates about the possible effects of cohesion on the test taker's performance mainly based on the previous research. In other words, although cohesion influences reader's text understanding, it is yet to be confirmed whether the cohesion of the reading passages has a significant impact on the test taker's actual score. In addition, the different cohesion between the tests might be the result of the different constructs of the tests. That is, the tests would differ from each other regarding what they are supposed to measure, thereby leading to the variation in text cohesion. Moreover, this study did not consult the actual processes in which the reading passages on CSAT, TEPS, EIKEN and TOEFL are developed. Information on the passage development procedures would help better understand the findings of this study.

Future research can extend the current study by using qualitative approaches in order to fully understand the cohesion of the passages on reading assessment. Moreover, experimental research might be undertaken to find out the effects of cohesion on the test taker's performance on English reading

assessment while considering their background knowledge. In addition, face-to-face interview, think-aloud, or eye-tracking method would provide information on the actual processes the test takers undergo. Future study might also compare the cohesion of the reading passages with other linguistic features and investigate the relations between them. Lastly, additional research can examine the interactions between the cohesion of the reading passages and the types of test items since cohesion would have different effects depending on the task types.

REFERENCES

- Ahmed, S. (2009). Methods in Sample Surveys: Simple Random Sampling Systematic Sampling. Retrieved November 11, 2015 from <http://ocw.jhsph.edu/courses/statmethodsforamplesurveys/pdfs/lecture2.pdf>
- Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University.
- Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York, NY: Routledge.
- Armbruster, B. B., & Anderson, T. H. (1985). Producing 'considerate' expository text: Or easy reading is damned hard writing. *Journal of Curriculum Studies*, 17(3), 247-274.
- Beck, I. L., McKeown, M. G., Omanson, R. C., & Pople, M. T. (1984). Improving the comprehensibility of stories: The effects of revisions that improve coherence. *Reading Research Quarterly*, 263-277.
- Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, 251-276.
- Best, R. M., Floyd, R. G., & Mcnamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts.

- Reading Psychology*, 29(2), 137-164.
- Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal*, 89(1), 37-53.
- Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83(3), 329.
- Bruss, M., Albers, M. J., & McNamera, D. (2004). *Changes in scientific articles over two hundred years: A coh-metrix analysis*. Paper presented at the Proceedings of the 22nd annual international conference on Design of communication: The engineering of quality documentation.
- Carrell, P. L. (1982). Cohesion Is Not Coherence. *TESOL quarterly*, 16(4), 479-488.
- Carrell, P. L. (1983). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a foreign language*, 1(2), 81-92.
- Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. *TESOL quarterly*, 441-469.
- Clarke, M. A. (1980). The short circuit hypothesis of ESL reading—or when language competence interferes with reading performance. *The Modern Language Journal*, 64(2), 203-209.

- Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology*, 22(1), 39-52.
- Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15-30.
- Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of semantic relations in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics*, 7, 55-74.
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38, 78-103.
- Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). *Using latent semantic analysis to improve access to textual information*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.
- Duran, N. D., Bellissens, C., Taylor, R. S., & McNamara, D. S. (2007). *Quantifying text difficulty with automated indices of cohesion and semantics*. Paper presented at the Proceedings of the 29th annual meeting of the cognitive science society.

- Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, 39(2), 212-223.
- Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3), 285-307.
- Freedle, R. (1997). The relevance of multiple-choice reading test data in studying expository passage comprehension: The saga of a 15 year effort towards an experimental/correlational merger. *Discourse Processes*, 23(3), 399-440.
- Frost, R. (2005). Orthographic systems and skilled word recognition. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 272-295). Malden, MA: Blackwell.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.
- Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371-398.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234.

- Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking reading comprehension*, 82-98.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2), 193-202.
- Grasser, A., & Hu, X. (2012). Moving forward on reading assessment. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 153-158). Maryland: Rowman & Littlefield Education.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. London: Longman Group Limited.
- Haswell, R. H. (1988). Critique: Length of text and the measurement of cohesion. *Research in the Teaching of English*, 428-433.
- Jalilehvand, M. (2012). The effects of text length and picture on reading comprehension of Iranian EFL students. *Asian Social Science*, 8(3), p329.
- Jeon, M. (2011). A Corpus-based analysis of the continuity of the reading materials in middle school English 1 and 2 textbooks with Coh-Metrix. *The Journal of Linguistic Science*, 56, 201-218.
- Jeon, M., & Azevedo, R. (2007). *Analyzing human tutorial dialogues for cohesion and coherence during hypermedia learning of a complex*

- science topic*. Paper presented at the Proceedings of the 29th Annual Meeting of the Cognitive Science Society.
- Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 219-239.
- Keenan, J. (2012). Measure for Measure: Challenges in Assessing Reading. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 77-88). Maryland: Rowman & Littlefield Education.
- Keshavarz, M. H., Atai, M. R., & Ahmadi, H. (2007). Content Schemata, Linguistic Simplification, and EFL Readers' Comprehension and Recall. *Reading in a foreign language*, 19(1), 19-33.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2), 163.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*: Cambridge university press.
- Kintsch, W. (2012). Psychological Models of Reading Comprehension and Their Implications for Assessment. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21-38). Maryland: Rowman & Littlefield Education.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language testing*,

19(2), 193-220.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*: Psychology Press.

Leslie, L., & Caldwell, J. (2009). 19 Formal and Informal Measures of Reading Comprehension. *Handbook of research on reading comprehension*, 403.

Lightman, E., McCarthy, P., Dufty, D., & McNamara, D. (2007). *The structural organization of high school educational texts*. Paper presented at the Proceedings of the twentieth international Florida artificial intelligence research society conference.

Lightman, E. J., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007). *Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters*. Paper presented at the Proceedings of the 29th Annual Meeting of the Cognitive Science Society.

Louwerse, M. M., Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (2007). Symbolic or embodied representations: A case for symbol interdependency. *Handbook of latent semantic analysis*, 107-120.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). *Variation in language and cohesion across written and spoken registers*. Paper presented at the Proceedings of the 26th Annual Meeting of the Cognitive Science Society.

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y., &

- McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes*, 6, 46-77.
- McCarthy, P. M., Lightman, E. J., Dufty, D. F., & McNamara, D. S. (2006). *Using Coh-Metrix to assess distributions of cohesion and difficulty: An investigation of the structure of high-school textbooks*. Paper presented at the Proceedings of the 28th annual conference of the Cognitive Science Society, Vancouver, Canada.
- McNamara, D. S., Graesser, A., & Louwerse, M. (2012). Sources of text difficulty: Across genres and grades. *Measuring up: Advances in how we assess reading ability*, 89-116.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*: Cambridge University Press.
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1), 1-43.
- McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22(3), 247-288.
- McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010).

- Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4), 292-330.
- McNamara, D. S., Ozuru, Y., Graesser, A. C., & Louwerse, M. (2006). *Validating coh-metrix*. Paper presented at the Proceedings of the 28th annual conference of the cognitive science society.
- Mehrpour, S., & Riazi, A. (2004). The impact of text length on EFL students' reading comprehension. *Asian EFL Journal*, 6(3), 1-13.
- Mislevy, R., & Sabatini, J. P. (2012). How research on reading and research on assessment are transforming reading assessment (or if they aren't, how they ought to). *Measuring up: Advances in how we assess reading ability*, 119-134.
- Newsom, R. S., & Gaite, A. (1971). Prose learning: Effects of pretesting and reduction of passage length. *Psychological reports*, 28(1), 123-129.
- Oblinger, D., & Oblinger, J. (2005). Is it age or IT: First steps toward understanding the net generation. *Educating the net generation*, 2(1-2), 20.
- Oh, S.-Y. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL quarterly*, 69-96.
- Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question?

- Behavior Research Methods*, 40(4), 1001-1015.
- Rawson, K. A., & Kintsch, W. (2005). Rereading Effects Depend on Time of Test. *Journal of Educational Psychology*, 97(1), 70.
- Rothkopf, E. Z., & Billington, M. (1983). Passage length and recall with test size held constant: Effects of modality, pacing, and learning set. *Journal of verbal learning and verbal behavior*, 22(6), 667-681.
- Service, E. T. (2014a). *The Official Guide to the TOEFL Test* (Fourth ed.). Seoul: YBM.
- Service, E. T. (2014b). *Official TOEFL iBT Tests* (First ed.). Seoul: YBM.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language testing*, 1(2), 147-170.
- Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders*, 25(1), 33-50.
- Surber, J. R. (1992). The effect of test expectation, subject matter, and passage length on study tactics and retention. *Literacy Research and Instruction*, 31(3), 32-40.
- van den Broek, P. (2012). Individual and Developmental Differences in Reading Comprehension: Assessing Cognitive Processes and Outcomes. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 39-58). Maryland: Rowman & Littlefield

Education.

Xi, Y. (2010). Cohesion studies in the past 30 years: Development, application, and chaos. *Language Society and Culture*(31), 139-147.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning*, 44(2), 189-219.

Young, D. J. (1999). Linguistic simplification of SL reading material: Effective instructional practice? *Modern Language Journal*, 350-366.

Zakaluk, B. L., & Samuels, S. J. (1988). *Readability: Its Past, Present, and Future*. Newark: International Reading Association.

Appendix 1. CSAT corpus

School Year (classification)	Administration	Number of Reading Passages Selected	Average Word Count per Passage
2016 (mock)	September, 2015	25	148
2016 (mock)	June, 2015	25	
2015 (official)	November, 2015	25	
2015 (mock)	September, 2014	25	
2015 (mock)	September, 2014	25	
Total 125			

Appendix 2. TEPS corpus

Korean Title of the Books	Test Number	Number of Reading Passages Selected	Average Word Count per Passage
서울대 텡스관리위원회 공식 최신기출 1200제 문제집 1	1	14	87
	2	13	
	3	13	
서울대 텡스관리위원회 공식 최신기출 1200제 문제집 2	1	14	
	2	13	
	3	13	
서울대 텡스관리위원회 공식 최신기출 1200제 문제집 3	1	14	
	2	13	
	3	13	
Total 120			

Appendix 3. EIKEN corpus

Test Session	Grade	Number of Reading Passages Selected	Average Word Count per Passage
2015 summer	1	5	400
	P1	5	
	2	5	
2014 winter	1	5	
	P1	5	
	2	5	
2014 fall	1	5	
	P1	5	
	2	5	
2014 summer	1	5	
	P1	5	
	2	5	
2013 winter	1	5	
	P1	5	
	2	5	
		Total 75	

Appendix 4. TOEFL corpus

Source	Test Number	Number of Reading Passages Selected	Average Word Count per Passage
http://www.ets.org/toefl		8	695
The Official Guide to the TOEFL Test	1	3	
	2	3	
	3	3	
		6	
Official TOEFL iBT Tests	1	3	
	2	3	
	3	3	
	4	3	
	5	3	
http://etest.chosun.com	35	3	
	36	3	
	37	3	
	38	3	
	39	3	
Total 53			

국문초록

본 연구는 대학수학능력시험, 텡스(TEPS), 에이켄(EIKEN), 토플(TOEFL)의 읽기영역 지문의 정합성을 분석하고자 한다. 정합성은 텍스트 내의 의미들의 관계를 가리키며 참조어(reference), 연결사(connectives), 어휘 등에 의해 형성된다. 이러한 정합성의 측정은 텍스트를 보다 심층적으로 분석함으로써 단어 및 문장 난이도에 기반한 전통적인 텍스트 난이도 측정도구의 대안이 될 수 있다. 정합성이 높은 텍스트는 이해를 돕는 반면, 정합성이 낮은 텍스트는 독자들의 텍스트 이해를 저해할 가능성이 있다. 특히 배경지식이 부족한 독자들은 끊어진 의미상의 고리를 연결하는데 필요한 지식이 부족하기 때문에 정합성이 낮은 텍스트를 이해하기 어려울 것이다. 따라서 수험자의 배경지식이 시험의 중요한 측정요소가 아니라면, 읽기지문은 배경지식이 부족한 수험자의 텍스트 이해를 심각하게 저해하지 않을 정도의 충분한 정합성을 가지고 있어야 한다.

본 연구는 읽기지문의 정합성을 측정하기 위해 웹에 기반한 텍스트 분석도구인 코메트릭스(Coh-Metrix)를 사용하였다. 본 연구는 참조적 정합성(referential cohesion), 잠재적 의미 분석(Latent Semantic Analysis), 연결사, 인과적 정합성을 측정하는 11개의 코메트릭스 인덱스를 선택하였다. 분석대상 지문은 총 373개로, 대학수학능력시험 125개, 텡스 120, 에이켄 75개, 토플 53개로 구성되었다.

분석 결과 연결사 빈도(Incidence of All Connectives)를 제외한 모든 인덱스에서 시험간 유의미한 차이가 있었다. 예상대로 가장 긴 지문의 토플은 대부분의 측정치에서 가장 높은 값을 나타냈다. 그러나 다른 시험들에 대한 결과는 혼재된 양상을 보였다. 텡스 지문은 대부분 에이켄과 정합성이 비슷하거나 더 높게 나타났

다. 아울러 대학수학능력시험은 텡스에 비해 지문이 길지만 대부분의 인덱스에서 가장 낮은 값을 보였다. 에이켄은 대학수학능력시험보다 정합성이 높게 나타났다. 대학수학능력시험과 텡스는 에이켄과 토플에 비해 지문간 정합성 편차가 매우 큰 것으로 나타났다. 또한 기존 정보와 새로운 정보의 비율을 계산하는 LSA Given/New에서 시험들 간의 정합성 차이가 가장 크게 나타났고, 명사 및 어근 중첩을 측정하는 Noun/Stem Local Overlap 에서도 시험들간 큰 차이가 있었다. LSA 인덱스는 참조적 정합성과 비슷한 정도로 시험들간의 차이를 뚜렷하게 보여 주었다. 인과적 정합성에서도 시험들 간의 차이는 있었지만, 그 정도는 미미했다. 마지막으로, 결과와 본 연구의 함의가 논의된다.

주요어: 응집성, 영어독해평가, 표준화 영어시험, 대학수학능력시험,

텡스(Teps), 에이켄(EIKEN), 토플(TOEFL), 코메트릭스(Coh-Metrix)

학 번: 2003-23711